Remarks

Claims 1-7 and 20-45 are pending. Claim 6 has been amended. Claims 8-19 are canceled. Claims 22-45 were not entered. Claims 46-69 are newly added. Claim 6 was amended to correct the spelling of "occurring." Support for new claim 46 can be found throughout the application, and specifically on page 42, lines 1-4. New claims 47, 50, 53, and 56 find support at least in original claims 1 and 6. New claims 50, 53, and 56 also find support at least on page 36, lines 15-19 and 27-33. New claims 48, 51, 54, and 57 find support at least on page 36, lines 15-19 and 27-33. New claims 49, 52, 55, and 58 find support at least on page 37, lines 6-7. New claims 59-69 find support at least in Figures 11A, 11B, 11C, 11D, 11E, 11F, 11G, 14A, 19A, 24A, and 30A, respectively, and on page 15, lines 15-25 (claims 59-65), page 16, lines 17-21 (claim 66), page 18, lines 9-15 (claim 67), from page 20, line 29, to page 21, line 4 (claim 68), and page 23, lines 23-9 (claim 69).

A replacement section for the section "BRIEF DESCRIPTION OF THE DRAWINGS" is provided with this response. The only changes are to the description of Figure 11, where reference to the wrong figure numeral is made.

Rejections Under 35 U.S.C. § 112, First Paragraph

1. Claims 1-7 and 20-21 were rejected under 35 U.S.C. § 112, first paragraph, as allegedly failing to comply with the written description requirement. Applicants respectfully traverse this rejection.

Claims 1-7 were considered to lack adequate written description on the basis that the specification and claims allegedly do not adequately describe the genus comprising regulatable gene expression constructs comprising riboswitches that are activated by a trigger molecule and produce a signal upon activation and which constructs further comprise a control strand, an aptamer domain, and an expression platform domain comprising a regulated strand.

Applicants respectfully traverse, first on the grounds that multiple examples of the genus of riboswitches have been provided, and second on the grounds that the premise of this rejection is not consistent with the law of written description.

A. Regarding the first point, the specification is replete with examples of structural features and sequence relationships of riboswitches. Importantly, the specification provides description of the key structural features and sequence relationships necessary for the operation of riboswitches in general, and provides multiple specific examples of such. For example, the specification (page 104, lines 13-20) states:

> Riboswitches that have been discovered are responsible for sensing metabolites that are critical for fundamental biochemical processes including adenosylcobalamin (AdoCbl) (see Example 1), thiamine pyrophosphate (TPP) (see Example 2), flavin mononucleotide (FMN), S-adenosylmethionine (SAM) (see Example 7), lysine (see Example 5) , guanine (see Example 6), and adenine (see Example 8). Upon interaction with the appropriate small molecule ligand, riboswitch mRNAs undergo a structural reorganization that results in the modulation of genes that they encode.

In each of the examples mentioned above, a detailed description of the riboswitch activated by a trigger molecule is given, along with an explicit discussion of how a trigger molecule interacts with the riboswitch. Thus, Applicants have described the general structure and operation of riboswitches; have identified the component parts of riboswitches, how they interconnect and operate, and how they can be recombined to form other riboswitches; and have provided a number of examples of riboswitches spanning a variety of genes and trigger molecules, thus solidifying both the validity of the general description and providing a representative number of examples of the structure of riboswitches. The numerous examples and consensus sequences provided clearly demonstrate Applicants' possession of the broad general subject matter of the present claims. It is hard to imagine how an applicant could provide more descriptive information of a pioneering invention than Applicants have provided.

The Office Action alleges that the specification and claims do not adequately describe the concise structural features (e.g., polynucleotide sequences, structures of all component parts of the gene expression constructs) that distinguish structures within the broadly claimed genus from those without. The Office Action states that the specification teaches the 5'-UTR of the B. subtilis xpt-pbuX mRNA as a guanine-responsive riboswitch. The Office Action goes on to state that the specification also teaches a comparison between the 5'-UTR fragment (of 185 nucleotides) and other bacterial sequences, whereby a conserved RNA motif, termed a "G box" has been identified as a domain for guanine riboswitches. The Office Action then alleges that conserved secondary and tertiary structures are likely a pre-requisite for adopting the required, yet undefined three-dimensional fold necessary for riboswitch function. The Office Action thus implies that a detailed secondary and tertiary structure must be described in the application. This is not the case.

First, Applicants note that riboswitches are made up of different domains that have different roles to play in the operation of the riboswitch. As fully described in the specification, riboswitches include an aptamer domain and an expression platform. The expression platform of riboswitches generally involves alternative stem structures. The principles of operation and application of platform domains are described in the specification. The formation of hybridized stem structures in RNA is well known in the art, and the examples and principles of the structure and operation of platform domains of riboswitches is described in the specification. The structure-function relationship of the stem structures of platform domains is thoroughly described in the specification and provides all that is required by the written description requirement for this element of the claims.

Aptamer domains are essentially RNA aptamers. RNA aptamers in other contexts have been known and described for many years. The aptamer domains of riboswitches bind to trigger molecules and communicate through the RNA strand to the platform domain. Applicants submit that aptamers can be used and applied in riboswitches based on the description provided in the specification. Applicants discovered that aptamers in riboswitches are modular and can be used and interchanged between riboswitches. As noted in the specification, the aptamer domain of the riboswitch readily adopts the required structure without interference from, and independent of, the other control structures of riboswitches, even in aptamer domains synthesized *in vitro*:

> These conclusions are drawn from the observation that aptamer domains synthesized in vitro bind the appropriate ligand in the absence of the expression platform (see Examples 2, 3 and 6). Moreover, structural probing investigations suggest that the aptamer domain of most riboswitches adopts a particular secondary- and tertiary-structure fold when examined independently, that is essentially identical to the aptamer structure when examined in the context of the entire 5′ leader RNA. This implies that, in many cases, the aptamer domain is a modular unit that folds independently of the expression platform (see Examples 2, 3 and 6).

Specification, page 30, lines 23-30.

Therefore, the generic primary and secondary structural features of riboswitches described in the specification produce the necessary three-dimensional structure, without the need for guidance or further description. This is borne out by Applicants' description and analysis of guanine-responsive riboswitches and their structure. Having identified an example of an aptamer in a guanine-responsive riboswitch (where the aptamer binds to guanine and related compounds), Applicants searched for, found, and identified consensus elements of other guanine aptamers in guanine-responsive riboswitches in other genes (see part C of Figure 41). The conservation and similarity of the primary sequence of aptamers in these riboswitches is strongly indicative that the higher level structure and aptamer function follow from the primary structure. Those of skill in the art would have been able to readily produce such functional riboswitches without concern for the three dimensional structure of the aptamer domain, because the three dimensional structure would have naturally folded into the correct orientation for functionality. Furthermore, as noted in the passage above, the aptamer domain can be a modular component that can be exchanged with other control sequences of the riboswitch. Because the aptamer domain can be exchanged with other control sequences, the riboswitch can comprise any aptamer. The specification comprises multiple examples of such aptamers (see, for example, Figures 11 and 41). Furthermore, aptamers in general are well known in the art and can be produced by known techniques, and are useful with the riboswitches disclosed in the specification.

The Advisory Action mailed February 8, 2008 merely repeats the flawed reasoning of the Office Action (merely referring to the detailed structures of example riboswitches and then alleging without support that such detail is required for an adequate written description) and provides not explanation of why such reasoning could support the present rejection in view of Applicants' arguments above. For the reasons discuss above and elsewhere herein, Applicants submit the claimed gene expression constructs are adequately described.

B. Subsequent to Applicants' invention, it has been confirmed that the consensus primary and secondary structural elements described in the present application naturally produce the structure required for riboswitch function. Tertiary structures of five classes of riboswitches have been solved and published (guanine-, adenine-, TPP-, SAM-, and glucosamine-6-phosphate-responsive riboswitches). In each publication the authors note how well Applicants' models and probing data (which corresponds to the models and data described in the present application) fit with the tertiary structures. For example, Serganov et al., Chem. Biol. 11:1729-1741 (2004), a copy of which is submitted with this Response, describes the crystal structure of add adenine-responsive riboswitch and the xpt guanine-responsive riboswitch. The add and xpt riboswitches are examples of riboswitches in the present application (see, for example, Figures 11E, 11F, 23, 24, 25 28, 35 and Examples 6 and 8). Serganov (2004) compares the crystal structure, and the functional significance of the structure revealed, with the conserved primary and secondary structural elements that characterize the adenine and guanine riboswitches. See, for example, Serganov et al. (2004) page 1737, second column, third and fourth paragraphs; page 1738, fist column, first, second and third paragraphs; and page 1738, second column, first paragraph. Serganov et al. (2004) confirms and concludes that the primary and secondary structural information and the conserved elements of adenine- and guanine-responsive riboswitches that were earlier identified have structural and functional significance in the crystal structure. For comparison to the crystal structure, a number of these passages in Serganov et al. (2004) refer to the primary and secondary structural information of citation number 24, Mandal & Breaker, Nature Struct. Mol. Biol. 11(1):29-35 (2004) (a copy of which is submitted with this Response) which describes some of that same structural information in the present application. For example, present Example 8 describes work reported in Mandal & Breaker and Figures 35-40 in the present application correspond to Figures 1-6, respectively, in Mandal & Breaker. Thus, Serganov et al. (2004) provides evidence that the conserved and consensus structural elements identified in the present application are significant in determining the crystal structure of the riboswitch.

Serganov et al., Nature 441:1167-1171 (2006), a copy of which is submitted with this Response, describes the crystal structure of the thiM thiamine pyrophosphate (TPP) responsive riboswitch. The thiM TPP-responsive riboswitch is one of the example riboswitches in the present application (see, for example, Figures 6B, 9A, 11B, 13A, and 13B and Example 2). Serganov (2006) compares the crystal structure, and the functional significance of the structure revealed, with the conserved primary and secondary structural elements that characterize the TPP riboswitches. See, for example, Serganov et al. (2006) page 1167, first column, last paragraph; page 1168, second column first paragraph; page 1168, second column, last paragraph; and page 1169, first column, third paragraph. Serganov et al. confirms and concludes that the primary and secondary structural information and the conserved elements of TPP-responsive riboswitches that were earlier identified have structural and functional significance in the crystal structure. For comparison to the crystal structure, a number of these passages in Serganov et al. (2006) refer to the primary and secondary structural information of citation number 4, Winkler et al., Nature 419:952-956 (2002) (of record), which describes some of that same structural information in the present application. For example, present Example 2 describes work reported in Winkler et al. and Figures 6, 7, 8, 9A-C, and 13B in the present application correspond to Figures 1-5, respectively, in Winkler et al. Thus, Serganov et al. (2006) provides evidence that the conserved and consensus structural elements identified in the present application are significant in determining the crystal structure of the riboswitch.

A new class of riboswitch has also been identified based on the consensus primary and secondary structural elements described in the present application, which confirms that riboswitch function predictably follows from the primary and secondary structural characteristics of the RNA. Kim et al., Proc. Natl. Acad. Sci. 104:16092-16097 (2007), a copy of which is submitted with this Response, describes the recently discovered 2'-deoxyguanosine-responsive riboswitch. The 2'-dG riboswitch was identified by searching sequences for primary and secondary structural elements based on the consensus structural elements of the guanine-responsive riboswitches that are described in the present application. In particular, Kim et al. refers to prior work with guanine- and adenine-responsive riboswitches as providing the basis of the identification of the new type of riboswitch, citing, for example, Mandal et al., Cell 113:577-586 (2003) (reference 26) (of record) and Mandal & Breaker, Nature Struct. Mol. Biol. 11(1):29-35 (2004) (reference 27). Example 6 in the present application describes work reported in Mandal et al. (2003) and Figures 23-29 in the present application correspond to Figures 1-7, respectively, in Mandal et al. (2003). As discussed above, present Example 8 describes work reported in Mandal & Breaker and Figures 35-40 in the present application correspond to Figures 1-6, respectively, in Mandal & Breaker. Once the 2'-dG riboswitch was identified in Kim et al. by these sequence and predicted secondary structural features, an example of a 2'-dG riboswitch was tested and determined to have the expected secondary structure and the expected riboswitch activity. Thus, Kim et al. also provides evidence that the consensus structural information provided in the present application represents structural information that adequately distinguishes the claimed riboswitches from other, non-riboswitch molecules. Further, the consistency of Applicants' description and understanding of riboswitch structure and function in the present application to subsequent findings regarding, and examples of, riboswitches is clear evidence that Applicants' were in possession of the riboswitches as presently claimed.

Applicants have also continued to identify additional riboswitches based on the original description and features described in the present application. Examples of such continued work include Barrick and Breaker, Genome Biology 8:R239 (2007), and Weinberg et al., Nucleic Acids Research 35(14):4809-4819 (2007, copies of which are submitted with this Response. These publications show that the general description of a new class of regulatory element based in RNA provided in the present application identified and described all of the key features and functions of riboswitches such that riboswitches could be identified and distinguished from what came before. Blount & Breaker, Nature Biotechnology 24(12):1558-1564 (2006), and Tucker & Breaker, Current Opinion in Structural Biology 15:342-348 (2005), copies of which are submitted with this Response, are reviews that describe riboswitch structure and function. These publications show the consistency between the description provided in the present application and the continued work in riboswitches since Applicants' invention.

C. As mentioned above, the premise of the arguments in the Office Action is not consistent with the law of written description. Compliance with the written description requirement need not involve the specific disclosure of every permutation of an invention, but should be commensurate with knowledge that comprises the state of the art. For example, in

Capon v. Eshhar v. Dudas, 76 USPQ2d 1078, 1082 (Fed. Cir. 2005), the court held that neither a complete nucleotide description nor operability of every permutation within a generally operable invention is required in order for an adequate written description of generically claimed nucleic acid constructs. Capon involved claims broadly drawn to nucleic acid constructs encoding a chimera of single-chain variable portions of antibodies and transmembrane lymphocyte signaling proteins. Both parties in an interference proceeding had appealed a decision by the Board of Patent Appeals and Interferences ("Board") that their specifications failed to provide an adequate written description of the claimed constructs. In particular, the Board stated that it could not be known whether all the permutations and combinations covered by the claims would be effective for the intended purpose, and that the claims were too broad because they might include inoperative species. Specifically, the Board stated that the disclosure of specific examples provided in each party's specification, in the absence of any sequence information within the specification, did not provide adequate written descriptive support for the invention. In reversing the Board's decision, the court in Capon held that since specific examples of the production of specified chimeric genes were provided in the specification, it was not necessary that every permutation within a generally operable invention be effective in order for an inventor to obtain a generic claim. The court also confirmed its long-standing precedent that the disclosure required to meet the written description requirement will vary with the nature and scope of the invention. In sum, the court in Capon concluded that knowledge in the art of the sequences of the nucleic acids that were joined to construct the chimeric DNA molecules, together with Appellants' disclosures of known methods for joining nucleic acid molecules to form chimeric DNAs, provided adequate written description of DNA molecules encoding chimeric receptors, and therefore recitation of exact nucleotide sequences was not required.

Applicants assert that this same logic applies to the claimed riboswitch constructs. Applicants have described the general structure and operation of riboswitches; have identified the component parts of riboswitches, how they interconnect and operate, how they can be recombined to form other riboswitches; and have provided a number of examples of riboswitches spanning a variety of genes and trigger molecules. Like the applicants' disclosures in Capon, Applicants have provided specific examples of riboswitches, as well as clear guidance of how to select the modular components thereof, such as the aptamer. Like the components of the chimeric DNAs in Capon, extensive sequence information is available for the claimed riboswitch components.[1]

As with the Board's basis for alleged unpatentability in Capon, the present rejection is based on an allegation that some of the presented riboswitch sequences would not be functional. As stated by the court in Capon, it is not necessary that every permutation be effective in order for an inventor to obtain a generic claim, provided that the effect was sufficiently demonstrated to characterize the invention. This principle applies to the present invention and facts with equal force and effect. Applicants submit that riboswitches were well demonstrated in the specification, as evidenced by the lengthy discussion of riboswitches therein (see page 104, lines 13-20, and the examples referenced in this passage, for example). Applicants have provided a variety of example riboswitches (and riboswitch components) and have identified consensus sequences for a number of riboswitches, which clearly qualifies as concise structural features that define the riboswitches and their components.

The Examiner alleges that the examples given, and the generic description of riboswitches, comprising an aptamer domain and an expression platform, the generic descriptions of structure function relationships for some identified (and proposed) stem structures of platform domains, and the sequence comparisons between previously described riboswitches found in nature, and sequence databases, together do not provide the concise structural features required for the very broad genus of compounds claimed. Applicants first note again that, as in Capon, the recitation of exact nucleotide sequences is not required for every permutation and combination of the claimed constructs. Applicants also note that, as discussed extensively above, the present specification provides a significant amount of information, including both specific and generic sequences, for myriad riboswitches. It is not seen how this fails to provide the "concise structural features" required by the rejection. On the contrary, the

[1] The fact that the relevant sequences in Capon were provided in the art rather in the specifications while Applicants here provide the riboswitch sequence information in the specification is not a relevant distinction. In fact, the inclusion of this sequence information in the present specification is more favorable for written description than was the case in Capon.

rejection merely concludes without evidence or sufficient reasoning that the extensive structural information provided is inadequate.[2] It can only be concluded that the rejection is applying a per se requirement of a type rejected by the court in Capon for a certain quality of written description. Such a per se and unsupported requirement is not supported by either the statute or the caselaw. Applicants have provided a full and complete disclosure, commensurate with knowledge that comprises the state of the art. One of skill in the art would have been able to identify riboswitches, and the components thereof, needed to make the claimed constructs, based on the disclosure in the specification.

Furthermore, in Falkner v. Inglis, 79 USPQ2d 1001 (Fed. Cir. 2006), the Federal Circuit found that Applicant need not spell out every detail of an invention, but only enough to convince a person of skill in the art that the inventor possessed the invention. At issue in Falkner were claims to a poxvirus lacking essential genes, for use as a vaccine. Although the specification at issue in Falkner neither identified, nor provided the sequence of, any essential poxvirus gene, essential regions of poxvirus were known in the art. The court, upholding a Board decision, found that the claims were adequately described. In support of its decision, the court held that:

> (1) examples are not necessary to support the adequacy of a written description (2) the written description standard may be met (as it is here) even where actual reduction to practice of an invention is absent; and (3) there is no *per se* rule that an adequate written description of an invention that involves a biological macromolecule must contain a recitation of known structure.

Although the specification at issue in Falkner neither identified, nor provided the sequence of, any essential poxvirus gene, essential regions of poxvirus were known in the art. The court agreed that those of skill in the art could easily select essential poxvirus genes. This is significant for the present rejection. The rejection dismisses Applicants' argument that those of

[2] The rejection seems to dismiss the majority of Applicants' disclosure of multiple riboswitch sequences and consensus sequences on the apparent basis that the Examiner does not believe that the identified sequences will function as riboswitches. This allegation does not support the present rejection for a number of reasons. First, there is no basis for the Examiner's assertion and it thus has no weight for the question of written description. Second, Applicants have stated that such sequences represent functional riboswitches, and such statements are to be believed unless there is sufficient evidence or reasoning to doubt them. In fact, there are good reasons why Applicants identification of riboswitch sequences is credible and creditable. In particular, the conservation of these sequences at locations proximate to transcription units and coding regions strongly supports that these sequences have functional significance.

skill in the art could readily produce the claimed constructs based on Applicants disclosure. The rejection discounts this argument on the basis that "Applicant must be in possession at the time of filing of an adequate representation of species for the broad genus of compounds claimed, not merely have the ability to screen for such peptides." However, it is clear from Falkner that "possession" for written description purposes does not require actual possession (i.e., reduction to practice) nor even a structural description of all elements of an invention. The specification at issue in Falkner did not describe any "essential genes" of any poxvirus, nor provide specific guidance for which genes of poxvirus were or were not essential. Nevertheless, the court in Falkner held that such specific description was not required to satisfy the written description requirement. Significantly, the court recognized that the fact that those of skill in the art could identify essential poxvirus genes (an identification found nowhere in the specification at issue in Falkner) was sufficient to satisfy the written description requirement. This is analogous to the present constructs where those of skill in the art could easily identify the claimed riboswitches by reference to Applicants' extensive disclosure. As a result, the present application satisfies the written description requirement for the present claims. For at least these reasons, the present rejections should be withdrawn.

2. Claims 1-7 and 20-21 were rejected under 35 U.S.C. § 112, first paragraph, because allegedly the specification, while being enabling for a method of searching for candidates of the genus comprising RNA comprising any riboswitch operably linked to a coding region, which riboswitch regulates expression of the RNA, and which riboswitch and coding region are heterologous to each other, and which riboswitch comprises an aptamer domain, a control strand and an expression platform domain comprising a regulated strand, and which regulated and/or control strands form a stem structure, and which riboswitch is optionally derived from a naturally occurring guanine-responsive riboswitch, and which riboswitch is activated by a trigger molecule and produces a signal upon activation by the trigger molecule, does not reasonably provide enablement for predictably making and designing the members of the broad genus of molecules claimed. Applicants respectfully traverse this rejection.

The Office Action maintains the argument that the "Applicant has not provided guidance in the specification toward a method of making and using a representative number of species of the expansive genus of molecules claimed." First of all, Applicants point out that the standard of "a representative number of species of the expansive genus of molecules claimed" differs from, and is insupportably stringent compared to, the true legal standard for enablement. The enablement requirement of 35 U.S.C. 112, first paragraph, is separate and distinct from the description requirement. Vas-Cath, Inc. v. Mahurkar, 935 F.2d 1555, 1563, 19 USPQ2d 1111, 1116-17 (Fed. Cir. 1991), MPEP Section 2161.

The MPEP (Section 2164.02) states that, "For a claimed genus, representative examples together with a statement applicable to the genus as a whole will ordinarily be sufficient if one skilled in the art (in view of level of skill, state of the art and the information in the specification) would expect the claimed genus could be used in that manner without undue experimentation. Proof of enablement will be required for other members of the claimed genus only where adequate reasons are advanced by the examiner to establish that a person skilled in the art could not use the genus as a whole without undue experimentation." In this case, the Examiner has not set forth sufficient reasons why the multiple and explicit examples given in the specification would not be applicable to the genus as a whole.

As previously stated, the specification (page 104, lines 13-20) states:

> Riboswitches that have been discovered are responsible for sensing metabolites that are critical for fundamental biochemical processes including adenosylcobalamin (AdoCbl) (see Example 1), thiamine pyrophosphate (TPP) (see Example 2), flavin mononucleotide (FMN), S-adenosylmethionine (SAM) (see Example 7), lysine (see Example 5) , guanine (see Example 6), and adenine (see Example 8). Upon interaction with the appropriate small molecule ligand, riboswitch mRNAs undergo a structural reorganization that results in the modulation of genes that they encode.

In each of the examples mentioned above, a detailed description of the riboswitch activated by a trigger molecule is given, along with an explicit discussion of how a trigger molecule interacts with the riboswitch. Thus, Applicants have described the general structure and operation of riboswitches; have identified the component parts of riboswitches, how they interconnect and operate, and how they can be recombined to form other riboswitches; and have provided a number of examples of riboswitches spanning a variety of genes and trigger molecules, thus solidifying both the validity of the general description and providing a representative number of examples of the structure of riboswitches. The numerous examples and consensus sequences provided clearly demonstrate Applicants have clearly provided representative examples of the genus of riboswitches and their components, together with a statement applicable to the genus as a whole, which is all the law requires.

The Office Action also alleges that "the ability to test various sequences for their ability to cleave target nucleic acid strands in the presence of various ligands, and the postulation of required, yet undefined structural constraints for riboswitch activities is not representative of the ability to predictably make and use the broad genus of compounds claimed." The Office Action goes on to allege that the specification fails to provide any particular guidance which resolves the known unpredictability in the art associated with determining the necessary sequences and structural components for designing functional riboswitches. The Office Action does not acknowledge Applicant's previous statement that the generic primary and secondary structural features of riboswitches described in the specification produce the necessary three-dimensional structure, without the need for further guidance. Those of skill in the art would have been able to readily produce such functional riboswitches without concern for the three dimensional structure of such, because the three dimensional structure would have naturally folded into the correct orientation for functionality.

Applicants would like to point out that in order to make a rejection, the examiner has the initial burden to establish a reasonable basis to question the enablement provided for the claimed invention. In re Wright, 999 F.2d 1557, 1562, 27 USPQ2d 1510, 1513 (Fed. Cir. 1993) (examiner must provide a reasonable explanation as to why the scope of protection provided by a claim is not adequately enabled by the disclosure). A specification disclosure which contains a teaching of the manner and process of making and using an invention in terms which correspond in scope to those used in describing and defining the subject matter sought to be patented must be taken as being in compliance with the enablement requirement of 35 U.S.C. 112, first paragraph, unless there is a reason to doubt the objective truth of the statements contained therein which must be relied on for enabling support (MPEP 2164.04). The Examiner has not given any reason why the objective truth of statements given by Applicant are doubted.

The rejection fails to provide any evidence or convincing arguments why those of skill in the art would have difficulty following the guidance provided by the present specification. Rather, the rejection consists only of conclusory statements unsupported by evidence or logical rationale. In particular, the rejection makes the following unsupported conclusions: "it would require undue experimentation beyond that taught in the instant specification, to produce the broad genus of compounds claimed," "Applicant has not provided guidance in the specification toward a method of making and using a representative number of species of the expansive genus of molecules claimed," "The specification as filed fails to provide any particular guidance which resolves the known unpredictability in the art associated with determining the necessary sequence and structural components for designing functional riboswitches[3] encompassed by the very broad genus claimed," "the invention as claimed would require de novo determination of sequence and structural characteristics, by trial and error, based on the identification and characterization of a representative number of species of the genus of compounds claimed," and "The examples provided do not enable one to make and use the broad genus of compounds claimed without undue experimentation." These read as mere opinion of the examiner; not conclusions based on the law and facts. The only fact on which these conclusions rely is the alleged extreme breadth of the claims. However, broad claims do not per se lack enablement. In fact, if the embodiments of a broad claim can each be made without the need for undue experimentation, then the number of embodiments encompassed by the claims does not render the claims nonenabled. Thus, the basis of the present rejection does not lead to a proper conclusion that the present claims lack enablement.

The Advisory Action merely repeats the allegation that "the ability to test various sequences for their ability to cleave target nucleic acid strands in the presence of various ligands, and the postulation of required, yet undefined structural constraints for riboswitch activities is

[3] Applicants note that there is no "known" unpredictability regarding riboswitch design, and the rejection presents no evidence that there is sufficient unpredictability to result in the need for undue experimentation, especially in view of the extensive description and guidance presented in the specification.

not representative of the ability to predictably make and use the broad genus of compounds claimed" and fails to acknowledge or address Applicants arguments in the last Response. Applicants first note that screening for active molecules is not inconsistent with enablement and claims can be fully enabled even if screening is required. As one well-known example, Applicants note that the need to screen for antibodies that have a particular binding specificity does not cause claims to such antibodies to lack enablement. The question is whether experimentation is undue, not whether screening is required or whether the resulting structures may not be completely predictable. As stated elsewhere herein, Applicants have provided extensive description of riboswitches and riboswitch structure and that the function of the riboswitches is related to this structure. Neither the Office Action nor the Advisory Action provide any evidence or convincing reasoning why screening of riboswitches for activity would require undue experimentation. Rather, the Advisory Action merely concludes that screening results in a lack of enablement. This is not the state of the law of enablement.

The Office Action alleges that the quantity of experimentation required to practice the invention as claimed would require the de novo determination of sequence and structural characteristics, by trial and error, based on the identification and characterization of a representative number of species of the genus of compounds claimed, whereby riboswitches are identified, designed, and constructed. As previously stated, in assessing whether undue experimentation would be required to make and use the claimed constructs, it is important to focus on what would actually have to be done in order to make and use the constructs. The claimed constructs comprise a regulatable gene expression construct comprising a nucleic acid molecule encoding an RNA comprising a riboswitch operably linked to a coding region, wherein the riboswitch regulates expression of the RNA, wherein the riboswitch and coding region are heterologous. Thus, one wishing to make and use the claimed constructs need only (1) obtain a riboswitch and a coding region and (2) operably link the two.

Applicants submit that practice of none of these steps would require undue experimentation and that the specification supports this conclusion. First, producing an RNA construct was well within the ability of those of skill in the art at the time of the invention and to do so would not have required undue experimentation. Applicants give multiple examples in the specification of riboswitches that can be used with the claimed invention, and carefully outline the components thereof, and how they can be obtained and used (see, for example, page 35 line 23 through page 42 line 14 of the specification). Applicant gives detailed information regarding the domains of the riboswitch, including both the aptamer and the expression platform. Thus, producing a construct as claimed would not require undue experimentation. Second, using the constructs was within the ability of those of skill in the art at the time of the invention. Applicants submit that it would not have required undue experimentation for those of skill in the art to use such constructs, as constructs in general and their applications were not only well known in the art at the time of the invention, but clearly discussed in the specification.

For all of the above reasons, applicants submit that the present claims are fully enabled and that the present rejection does not provide persuasive evidence or argument to the contrary. Accordingly, the present rejection should be withdrawn.

Rejection Under 35 U.S.C. § 102

Claims 1-7 and 20 were rejected under 35 U.S.C. § 102(a) as being anticipated by Breaker et al. (Curr. Opin. Biotech 13:31-39, Feb. 1, 2002). Applicants respectfully traverse this rejection.

The claims are drawn to a regulatable gene expression construct comprising a nucleic acid molecule encoding an RNA comprising a riboswitch operably linked to a coding region, wherein the riboswitch regulates expression of the RNA, wherein the riboswitch and coding region are heterologous. Breaker et al. does not teach riboswitches at all, but instead teaches intramolecularly cleaving ribozymes. As disclosed in the specification, both an aptamer domain and an expression platform make up a riboswitch, as opposed to a ribozyme. Breaker et al. teaches no such elements. Specifically, Breaker et al. does not disclose riboswitches operably linked to a coding region.

The Advisory Action erroneously argues that the ribozyme of Breaker et al. constitutes a coding region as claimed. This is not the case. A coding region is a portion of nucleic acid that codes for amino acids. The ribozymes of Breaker et al. do not code for amino acids. Rather, the ribozyme itself has catalytic activity. This is a crucial error in the premise of the rejection. For at least this reason, Breaker et al. fails to disclose each of the claimed elements.

Furthermore, Breaker et al. do not teach or suggest that the disclosed constructs involve expression of RNA. The claims are specifically drawn to a riboswitch that regulates the expression of RNA, unlike the constructs of Breaker et al., which teach only constructs that have a catalytic function when bound by an effector. The various portions of Breaker et al. cited in the Office Action do not describe any riboswitch (or any genetic switch) that regulates expression of RNA as claimed. The text on pages 31 and 32 of Breaker et al. merely discuss catalytic RNAs, the possibility and usefulness of allosteric ribozymes if they can be produced, and the hope that allosteric ribozymes could be used in the future to make engineered constructs (the text on page 38 also discusses the hope that allosteric ribozymes could be used in the future to make engineered constructs). This is mere speculation is does not enable production or use of such constructs. The catalytic RNAs and allosteric ribozymes of Breaker et al. are not described in Breaker et al. as controlling RNA expression. Rather, Breaker et al. merely expresses the hope that such ribozymes could be used some day in genetic control systems. The text on page 38 of Breaker et al. specifically notes that this hoped for use "poses many challenges to ribozyme engineers." Figures 2 and 3 merely illustrate schemes for selecting and enriching active ribozymes. There is no control of RNA expression by the ribozymes in Figure 2 or 3. Accordingly, Breaker et al. fails to disclose at least two features of the claimed constructs: linkage of a riboswitch to a coding region and regulation of RNA expression by a riboswitch. Because Breaker et al. fails to disclose every limitation of the claimed constructs, Breaker et al. fails to anticipate the claims. Applicants therefore respectfully request withdrawal of this rejection.

Conclusion

Pursuant to the above amendments and remarks, reconsideration and allowance of the pending application is believed to be warranted. The Examiner is invited and encouraged to directly contact the undersigned if such contact may enhance the efficient prosecution of this application to issue.

A credit card payment in the amount of $1225.00, representing $820.00 for the fee for a small entity under 37 C.F.R. § 1.17(a)(4) and $405.00 for the fee for a small entity under 37 C.F.R. § 1.17(e), a Request for a Four Month Extension of Time, and a Request for Continued

Examination are enclosed. This amount is believed to be correct; however, the Commissioner is hereby authorized to charge any additional fees which may be required, or credit any overpayment to Deposit Account No. 14-0629.

Respectfully submitted,

NEEDLE & ROSENBERG, P.C.

/Janell T. Cleveland/

Janell T. Cleveland
Registration No. 53,848

NEEDLE & ROSENBERG, P.C.
Customer Number 23859
(678) 420-9300
(678) 420-9301 (fax)

<table>
<tr><th colspan="4">CERTIFICATE OF ELECTRONIC TRANSMISSION UNDER 37 C.F.R. § 1.8</th></tr>
<tr><td colspan="4">I hereby certify that this correspondence, including any items indicated as attached or included, is being transmitted via electronic transmission via EFS-Web on the date indicated below.</td></tr>
<tr><td>Name of Person Mailing</td><td colspan="3">Janell T. Cleveland</td></tr>
<tr><td>Signature</td><td>/Janell T. Cleveland/</td><td>Date</td><td>May 13, 2008</td></tr>
</table>

Research

Open Access

The distributions, mechanisms, and structures of metabolite-binding riboswitches

Jeffrey E Barrick[*†] and Ronald R Breaker[*‡§]

Addresses: [*]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520-8103, USA. [†]Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI48824-4320, USA. [‡]Howard Hughes Medical Institute, Yale University, New Haven, Connecticut 06520-8103, USA. [§]Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, Connecticut 06520-8103, USA.

Correspondence: Ronald R Breaker. Email: ronald.breaker@yale.edu

Published: 12 November 2007

Genome **Biology** 2007, **8:**R239 (doi:10.1186/gb-2007-8-11-r239)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2007/8/11/R239

Received: 26 July 2007
Revised: 1 October 2007
Accepted: 12 November 2007

Abstract

Background: Riboswitches are noncoding RNA structures that appropriately regulate genes in response to changing cellular conditions. The expression of many proteins involved in fundamental metabolic processes is controlled by riboswitches that sense relevant small molecule ligands. Metabolite-binding riboswitches that recognize adenosylcobalamin (AdoCbl), thiamin pyrophosphate (TPP), lysine, glycine, flavin mononucleotide (FMN), guanine, adenine, glucosamine-6-phosphate (GlcN6P), 7-aminoethyl 7-deazaguanine (preQ$_1$), and *S*-adenosylmethionine (SAM) have been reported.

Results: We have used covariance model searches to identify examples of ten widespread riboswitch classes in the genomes of organisms from all three domains of life. This data set rigorously defines the phylogenetic distributions of these riboswitch classes and reveals how their gene control mechanisms vary across different microbial groups. By examining the expanded aptamer sequence alignments resulting from these searches, we have also re-evaluated and refined their consensus secondary structures. Updated riboswitch structure models highlight additional RNA structure motifs, including an unusual double T-loop arrangement common to AdoCbl and FMN riboswitch aptamers, and incorporate new, sometimes noncanonical, base-base interactions predicted by a mutual information analysis.

Conclusion: Riboswitches are vital components of many genomes. The additional riboswitch variants and updated aptamer structure models reported here will improve future efforts to annotate these widespread regulatory RNAs in genomic sequences and inform ongoing structural biology efforts. There remain significant questions about what physiological and evolutionary forces influence the distributions and mechanisms of riboswitches and about what forms of regulation substitute for riboswitches that appear to be missing in certain lineages.

Background

Riboswitches are autonomous noncoding RNA elements that monitor the cellular environment and control gene expression [1-4]. More than a dozen classes of riboswitches that respond to changes in the concentrations of specific small molecule ligands ranging from amino acids to coenzymes are currently known. These metabolite-binding riboswitches are classified according to the architectures of their conserved aptamer domains, which fold into complex three-dimensional structures to serve as precise receptors for their target molecules. Riboswitches have been identified in the genomes of archaea, fungi, and plants; but most examples have been found in bacteria.

Regulation by riboswitches does not require any macromolecular factors other than an organism's basal gene expression machinery. Metabolite binding to riboswitch aptamers typically causes an allosteric rearrangement in nearby mRNA structures that results in a gene control response. For example, bacterial riboswitches located in the 5' untranslated regions (UTRs) of messenger RNAs can influence the formation of an intrinsic terminator hairpin that prematurely ends transcription or the formation of an RNA structure that blocks ribosome binding. Most riboswitches inhibit the production of unnecessary biosynthetic enzymes or transporters when a compound is already present at sufficient levels. However, some riboswitches activate the expression of salvage or degradation pathways when their target molecules are present in excess. Certain riboswitches also employ more sophisticated mechanisms involving self-cleavage [5], cooperative ligand binding [6], or tandem aptamer arrangements [7].

Many aspects of riboswitch regulation have not yet been critically and quantitatively surveyed. To forward this goal, we have compiled a comparative genomics data set from systematic database searches for representatives of ten metabolite-binding riboswitch classes (Table 1). The results define the overall taxonomic distributions of each riboswitch class and outline trends in the mechanisms of riboswitch-mediated gene control preferred by different bacterial groups. The expanded riboswitch sequence alignments resulting from these searches include newly identified variants that provide valuable information about their conserved aptamer structures. Using this information, we have re-evaluated the consensus secondary structure models of these ten riboswitch classes. The updated structures reveal that certain riboswitch aptamers utilize previously unrecognized examples of common RNA structure motifs as components of their conserved architectures. They also highlight new base-base interactions predicted with a procedure that estimates the statistical significance of mutual information scores between alignment columns.

Results and discussion

Riboswitch identification overview

Metabolite-binding riboswitch aptamers are typical of complex functional RNAs that must adopt precise three-dimensional shapes to perform their molecular functions. A conserved scaffold of base-paired helices organizes the overall fold of each aptamer. The identities of bases within most helices vary during evolution, but changes usually preserve base pairing to maintain the same architecture. In contrast, the base identities of nucleotides that directly contact the tar-

Table 1

Sources of riboswitch sequence alignments and molecular structures

		References		
Riboswitch class	Rfam accession	Seed alignment	Other alignments	Molecular structures
Thiamine pyrophosphate (TPP)	RF00059	[41]	[48]	[71-73]
Adenosylcobalamin (AdoCbl)	RF00174	[39]	[20]	
Lysine	RF00168	[37]	[21]	
Glycine	RF00504	[6]		
S-Adenosylmethionine class 1 (SAM-I)	RF00162	[94]	[9,52]	[78]
Flavin mononucleotide (FMN)	RF00050	[56]		
Guanine and adenine (purine)	RF00167	[22]		[95-97]
Glucosamine-6-phosphate (GlcN6P)	RF00234	[23]		[28,30]
7-Aminoethyl 7-deazaguanine (preQ$_1$)	RF00522	[40]		
S-Adenosylmethionine class 2 (SAM-II)	RF00521	[18]		

Riboswitches are named for the metabolite that they sense with standard abbreviations in parentheses. Rfam database numbers are provided for each riboswitch along with references to the seed alignments we used to train covariance models for database searches in this study, other published multiple sequence alignments, and three-dimensional molecular structures.

get molecule or stabilize tertiary interactions necessary to assemble a precise binding pocket are highly conserved even in distantly related organisms. Additionally, many riboswitches tolerate long nonconserved insertions at specific sites within their structures. These 'variable insertions' typically adopt stable RNA stem-loops that do not interfere with folding of the aptamer core.

Nearly all of the riboswitches discovered to date are *cis*-regulatory elements. For example, bacterial riboswitches are almost always located upstream of protein-coding genes related to the metabolism of their target molecules. Therefore, the genomic contexts of putative hits returned by an RNA homology search can be used to recognize legitimate riboswitches even when a search algorithm returns many false positives. Using this tactic, one can iteratively refine the description of a riboswitch aptamer by incorporating authentic low scoring hits into a new structure model and then re-searching the sequence database.

Several riboswitches were first identified as widespread RNA elements based on the presence of a highly conserved 'box' sequence within their structures. BLAST searches for the B12 box [8], S box [9], and THI box [10] sequences are effective for discovering many examples of the adenosylcobalamin (AdoCbl), *S*-adenosylmethionine (SAM)-I, and thiamin pyrophosphate (TPP) riboswitches, respectively. Other search techniques score how well a sequence matches a template of conserved bases and base-paired helices that the user manually devises from known examples of the riboswitch aptamer. The RNAmotif program performs this sort of generalized pattern matching [11]. A third strategy computationally defines and then searches for ungapped blocks of sequence conservation that are characteristic of a given riboswitch and spaced throughout its structure [12]. While these methods can be effective, they generally do not fully exploit the information contained in multiple sequence alignments of functional RNA families to efficiently identify highly diverged members.

Covariance models (CMs) are generalized probabilistic descriptions of RNA structures that offer several advantages over other homology search methods [13]. CMs can be directly trained on an input sequence alignment without time-consuming manual intervention. They also provide a more complete model of the sequence and structure conservation observed in functional RNA families that incorporates: first-order sequence consensus information; second-order covariation, where the probability of observing a base in one alignment column depends on the identity of the base in another column; insert states that allow variable-length insertions; and deletion states that allow omission of consensus nucleotides. This complexity comes at a computational cost, but several filtering techniques have recently been developed that make CM searches of large databases practical [14-16]. For example, CMs have been used to find divergent homologs of *Escherichia coli* 6S RNA [17] and define a variety of regulatory RNA motifs in α-proteobacteria [18]. The Rfam database [19] maintains hundreds of covariance models for identifying a wide variety of functional RNAs, including riboswitches.

In the present study, we used covariance models to systematically search for ten classes of metabolite-binding riboswitches in microbial genomes, environmental sequences, and selected eukaryotic organisms. The riboswitch sequence alignments used to train these CMs were derived from a variety of published and unpublished sources (Table 1). The genomic contexts of prospective riboswitch hits were examined to confirm that each was appropriately positioned to function as a regulatory element. In general, CMs trained on the input alignments were able to discriminate valid riboswitch sequences from false positive hits on the basis of CM scores alone. The most common exceptions were spuriously high-scoring AU-rich matches to the smaller riboswitch models (for example, the purine riboswitch) and *bona fide* low-scoring hits with variable insertions at unusual positions in the more structurally complex riboswitch classes.

Prospective riboswitch matches were also examined to ensure that they conformed to known aptamer structure constraints. In certain cases, it was necessary to manually correct portions of the automated sequence alignments defined by the maximally scoring path of each hit through the states of the CM. For example, CMs model only hierarchically nested base pairs for algorithmic speed [13]. Consequently, the pseudoknotted helices and pairings present in several riboswitches were aligned by hand to achieve the desired accuracy. The automated CM alignments also tend to incorrectly shift nucleotides when deletions of consensus positions result in ambiguity concerning the optimal placement of remaining sequences. The alignments of new RNA structure motifs and base-base interactions described later that were not present in the seed alignments used to train the covariance models were also manually adjusted. Multiple sequence alignments of the resulting curated riboswitch hits are available as Additional data files 1 and 2.

Riboswitch distributions

The phylogenetic distributions of the ten riboswitch classes were mapped from these search results (Figure 1). Members of the TPP riboswitch class are the only metabolite-binding RNAs known to occur outside of eubacteria. TPP riboswitch representatives are found in euryarchaeal, fungal, and plant species. The AdoCbl riboswitch is the most widespread class in bacteria, but TPP, flavin mononucleotide (FMN), and SAM-I riboswitches are also common in many groups. Glycine and lysine riboswitches have more fragmented distributions. They are widespread in certain bacterial groups, but appear to be missing from others. Finally, the glucosamine-6-phosphate (GlcN6P), purine, 7-aminoethyl 7-deazaguanine (preQ$_1$), and SAM-II riboswitches were identified in only a few groups of bacteria. Interestingly, the SAM-I and SAM-II

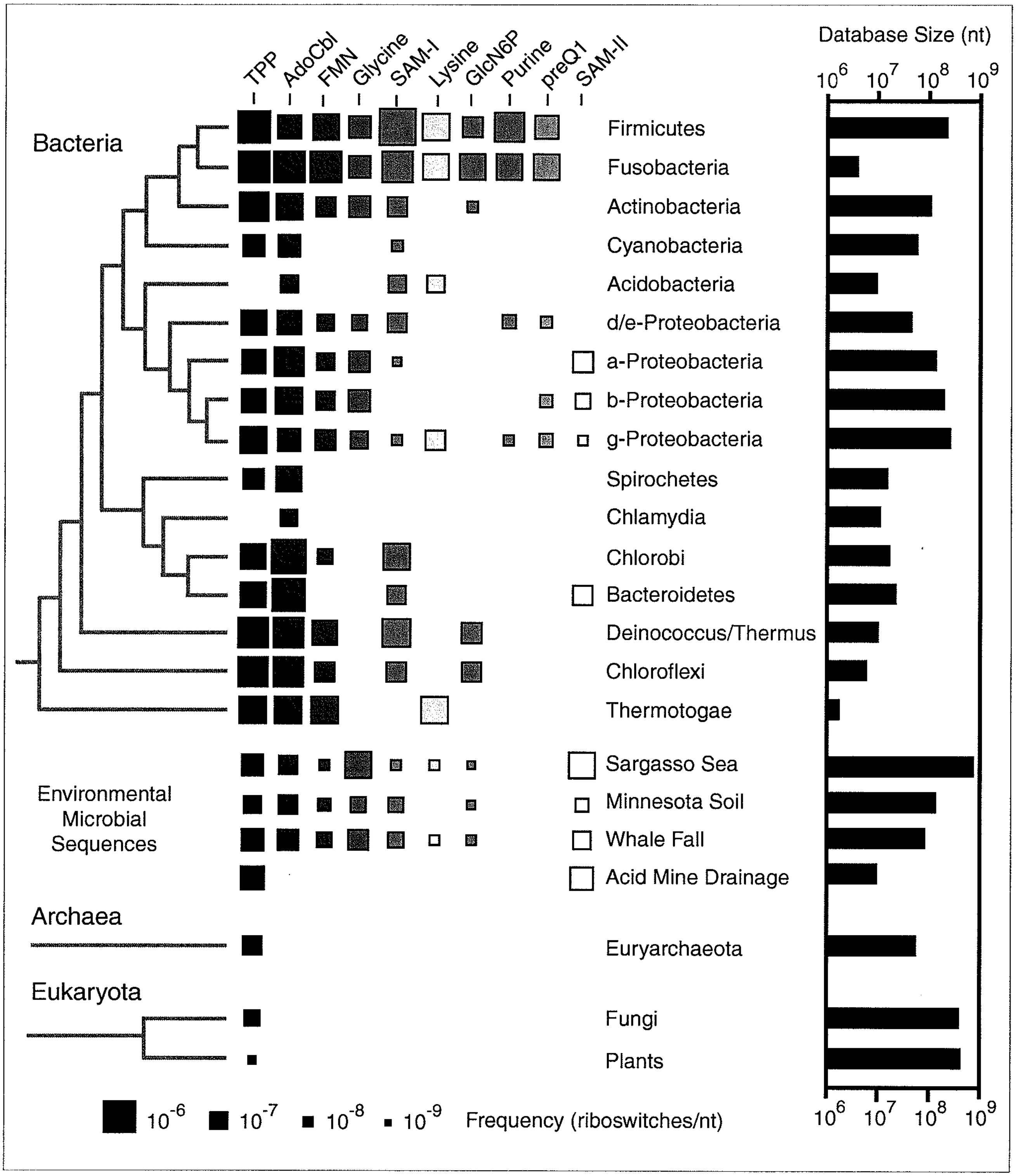


Figure 1

Riboswitch distributions. The dimensions of each square are proportional to the frequency with which a given riboswitch occurs in the corresponding taxonomic group. A phylogenetic tree with the standard accepted branching order for each group of organisms is shown on the left. For bacteria, this tree is adapted from [92] with the addition of Fusobacteria [93]. On the right is a graph depicting the total number of nucleotides from each taxonomic division in the sequence databases that were searched.

aptamer distributions overlap slightly. Examples of both SAM-sensing riboswitch classes were found in α-Proteobacteria, γ-Proteobacteria, and Bacteroidetes, but no single bacterial species was found to carry both SAM-I and SAM-II riboswitch classes.

It is possible that many of the relatively isolated examples where riboswitches occur only sporadically in certain clades (for example, SAM-I, SAM-II, purine, and preQ_1 in γ-Proteobacteria) may be examples of horizontal DNA transfer. There is some evidence that this process has been important for the dispersal of riboswitches into new bacterial genomes. Entire transcriptional units containing AdoCbl riboswitches and their associated biosynthetic operons appear to have been transferred from *Bacillus/Clostridium* species to enterobacteria at some point [20]. In contrast, no evidence of recent horizontal transfer was observed in phylogenetic trees of lysine riboswitch aptamers, despite their disjointed distribution across different taxonomic groups [21].

Firmicutes (low G+C Gram-positive bacteria) appear to make the most extensive use of the riboswitch classes examined in this study. Every riboswitch except SAM-II is widespread in this clade, and most aptamer classes occur multiple times per genome. For example, *Bacillus subtilis* carries at least 29 riboswitches (5 TPP, 1 AdoCbl, 2 FMN, 1 glycine, 11 SAM-I, 2 lysine, 1 GlcN6P, 4 guanine, 1 adenine, and 1 preQ_1) controlling approximately 73 genes. Experimental and computational efforts to identify riboswitches have been focused specifically on *B. subtilis* [22,23], so it is possible that the overrepresentation of these ten riboswitch classes in Firmicutes reflects a discovery bias. Indeed, new computational searches are beginning to identify riboswitch classes that are predominantly used by other groups of bacteria [18,24].

As a whole, γ-Proteobacteria employ a mixture of these ten riboswitch classes that is comparable to the diversity found in Firmicute species. However, individual species usually carry fewer riboswitch classes overall and fewer representatives of each class. For example, *E. coli* has six riboswitches (three TPP, one AdoCbl, one FMN, and one lysine) from the ten classes examined, which regulate a total of sixteen genes.

Deeply branched bacteria such as *Deinococcus/Thermus* and *Thermotoga* species also appear to utilize a variety of riboswitches. However, no riboswitch sequences have yet been identified in *Aquifex* species, and riboswitches also seem to occur only rarely in *Chlamydia* species, Cyanobacteria, and Spirochetes. However, the sequence database sizes for many of these bacterial groups are relatively small so the observed frequencies will probably need to be revised as more genomic sequences become available.

As expected, representatives of almost all ten riboswitch classes are found in sequences from shotgun cloning projects that target environments supporting diverse bacterial communities. These sources of additional sequences have been helpful in some cases for defining consensus structure models and adding statistical merit to mutual information calculations (see below). It is notable that glycine and SAM-II riboswitches are unusually common in Sargasso Sea metagenomic sequences [25]. This data set appears to be contaminated with some non-native *Shewanella* and *Burkholderia* sequences [26], but the large number of SAM-II matches probably accurately reflects the abundance of α-Proteobacteria in this environment.

Riboswitch mechanism overview

GlcN6P riboswitches are ribozymes that harness a self-cleavage event to repress expression of downstream *glmS* genes [5]. Members of this class are unique compared to other riboswitches because they adopt a preformed binding pocket for glucosamine-6-phosphate [27,28] and use the metabolite target as a cofactor to accelerate RNA cleavage [28-30]. The nine other riboswitch classes studied here utilize ligand-induced changes in 'expression platform' sequences to control a variety of gene expression processes [1]. The architectures of riboswitch expression platforms can be used to predict their gene control mechanisms on a genomic scale, as described below.

Riboswitches typically contain disordered regions in their conserved aptamer cores that become structured upon metabolite binding. These changes may trigger rearrangements in additional expression platform structures located outside of the aptamer, such that two alternative conformations with mutually exclusive base-paired architectures exist for the entire riboswitch. Some riboswitches operate at thermodynamic equilibrium [31]. They are able to interconvert between these ligand-bound and ligand-free structures in the context of the full-length RNA. Regulation by other riboswitches is kinetically controlled [32-35]. The relative speeds of transcription and co-transcriptional ligand binding dominate a one-time decision as to which folding pathway to follow. The active and inactive conformations of these riboswitches are trapped in the final RNA molecule and do not readily interconvert on a time scale that is relevant to the gene control system.

In most riboswitches, bases from the aptamer's outermost P1 'switching' helix, which is enforced in the ligand-bound conformation, pair to expression platform sequences to form an alternative structure in the absence of ligand, for example, [36,37]. However, some riboswitches harness shape changes elsewhere in their aptamers to regulate gene expression. AdoCbl riboswitches usually rely on the ligand-dependent formation of a pseudoknot between a specific C-rich loop and sequences outside the aptamer core to exert gene control [20,38,39]. SAM-II aptamers enforce a distal pseudoknot to interface with their expression platforms [18], and preQ_1 riboswitches sequester conserved 3' tail sequences upon metabolite binding [40].

Riboswitches can use ligand-induced structure changes to control gene expression in a variety of contexts. For example, the TPP riboswitches found in eukaryotes reside in introns located near the 5' ends of fungal pre-mRNAs [41-43] or in the 3' UTRs of plant pre-mRNAs [41]. Ligand binding modulates splicing of these introns, generating alternative-processed mRNAs that are expressed at different levels. In each example studied, a portion of the P4-P5 stem region pairs near a 5' splice-site, and this pairing is displaced when TPP is bound [43] (A Wachter, M Tunc-Ozdemir, BC Grove, PJ Green, DK Shintani, RRB, unpublished data). In contrast, almost all bacterial riboswitches occur in the 5' UTRs of mRNAs. Metabolite binding to these riboswitches generally regulates either transcription or translation of the encoded genes.

Bacterial riboswitches that regulate transcription usually control the formation of intrinsic terminator stems located within the same 5' UTR. Intrinsic terminators are stable GC-rich stem-loops followed by polyuridine tracts that cause RNA polymerase to stall and release the nascent RNA with some probability [44,45]. Certain glycine [6] adenine [46], and lysine [21] riboswitches with ON genetic logic use structural rearrangements triggered by metabolite binding to bury pieces of terminator stems in alternative pairing interactions. However, most riboswitches controlling transcription are OFF switches that add an extra folding element to reverse this logic. Metabolite binding to these riboswitches disrupts an antiterminator, which normally sequesters bases required to form the terminator stem, allowing the terminator to form and repress gene expression. Similar antiterminator/terminator trade-offs occur in bacterial RNAs regulated by protein- or ribosome-mediated transcription attenuation mechanisms [47].

Bacterial riboswitches that regulate translation typically use ligand-induced structure changes to block translation initiation. Unlike riboswitches with transcription control mechanisms, which require very specific terminator structures in their expression platforms, the RNA structures that prevent translation initiation may be more varied. Sometimes, they rely on simple hairpins that sequester the ribosome binding site (RBS) of the downstream gene in a base-paired helix. In these cases, a riboswitch with OFF genetic logic can harness metabolite binding to disrupt a mutually exclusive antisequestor pairing, allowing the sequestor hairpin to form and attenuate translation. More convoluted base-pairing trade-offs and shape changes may operate in other expression platforms to alter the efficiency of translation initiation in response to ligand binding.

Two variants of these mechanisms that dispense with or combine the elements of a typical bacterial riboswitch expression platform are worth noting. Some riboswitches bury the RBS of the downstream gene within their conserved aptamer cores [48,49]. Thus, ligand binding directly attenuates translation without the involvement of any additional expression platform sequences. Other riboswitches regulate the formation of a transcription terminator located so close to the adjacent open reading frame that its RBS resides within the 3' side of the terminator hairpin [48]. Riboswitches with these dual expression platforms could attenuate transcription and, if termination does not occur, could also inhibit translation.

Metabolite-dependent inhibition of ribosome binding has been proven *in vitro* for the *E. coli* AdoCbl riboswitch located upstream of the *btuB* gene [50]. In addition, *in vivo* expression assays using translational fusions between AdoCbl riboswitches and reporter genes indicate that control of translation is occurring [38]. However, other co- or post-transcription mechanisms might also contribute to the observed gene expression changes. For example, AdoCbl riboswitches from *E. coli* and *B. subtilis* can be cleaved by RNase P [51]. Such findings raise the interesting possibility that differential RNA processing or degradation caused by ligand-induced conformational changes might be the primary mechanism by which some riboswitches regulate gene expression.

There is one interesting instance where a *Clostridium acetobutylicum* SAM-I riboswitch appears to regulate protein expression through an antisense RNA intermediate [52]. This riboswitch is located immediately downstream, and in the opposite orientation from, an operon encoding a putative salvage pathway for converting methionine to cysteine. It has an expression platform, consisting of a typical terminator/antiterminator arrangement, with OFF genetic logic. Presumably, when SAM (and consequently methionine) pools are low, transcription of the full-length antisense RNA causes inhibition and degradation of the sense mRNA as is observed in some bacterial regulatory systems that employ small RNAs [53]. When SAM levels are high, the SAM-I riboswitch will prematurely terminate the antisense transcript, allowing expression of this operon to recycle excess methionine.

In some instances, riboswitches or their components are found in tandem arrangements. Almost all glycine riboswitches consist of two aptamers that regulate a single downstream expression platform [6]. In the genomic sequences searched here, 88% of the mRNA leaders containing one glycine aptamer also carry a second aptamer. Cooperative binding of two ligand molecules by these glycine riboswitches yields a genetic switch that is more 'digital', that is, more responsive to smaller changes in ligand concentration, than a single aptamer.

Far less common are tandem arrangements of other riboswitch classes such as TPP [7,54,55] or AdoCbl [55]. Fewer than 1% of the UTRs regulated by these riboswitch classes contain multiple aptamers. In these cases, each aptamer appears to function as an independent riboswitch that regulates its own expression platform to yield a more digital, compound genetic switch [7]. Also rare are tandem arrangements wherein representatives of two different riboswitches are in the same UTR. In the *metE* mRNA leader from *Bacillus clausii*, a SAM-I and an AdoCbl riboswitch independently control transcription termination to combinatorially regulate expression of this gene in response to two different metabolite inputs [55].

Riboswitch mechanisms

A decision tree was established for computationally classifying the gene control mechanisms of microbial riboswitches (Figure 2). The five categories assigned are: transcription attenuation; dual transcription and translation attenuation; translation attenuation; direct translation attenuation; and antisense regulation. The same mechanisms have been predicted for TPP [48], AdoCbl [20], FMN [56], and lysine [21] riboswitches in previous comparative studies. The use of the term attenuation here does not imply that a switch operates with OFF genetic logic, that is, gene expression may be attenuated in the ligand-free state and relieved by metabolite binding. Overall, computational assignments by this procedure have an accuracy of 88% when compared to expert predictions of TPP riboswitch mechanisms [48].

It is important to note that the decision tree does not explicitly predict RBS-hiding structures in expression platforms. Rather, it assumes that control of translation initiation is the most likely mechanism for riboswitches not classified into the other categories. It is possible that these riboswitches could operate by mechanisms other than the five assigned by this procedure (as described above). Another caveat is that this prediction scheme considers only intrinsic terminator structures consisting of RNA stem-loops followed by polyuridine tails. These are currently the only structures that riboswitches with transcription attenuation mechanisms are known to reg-

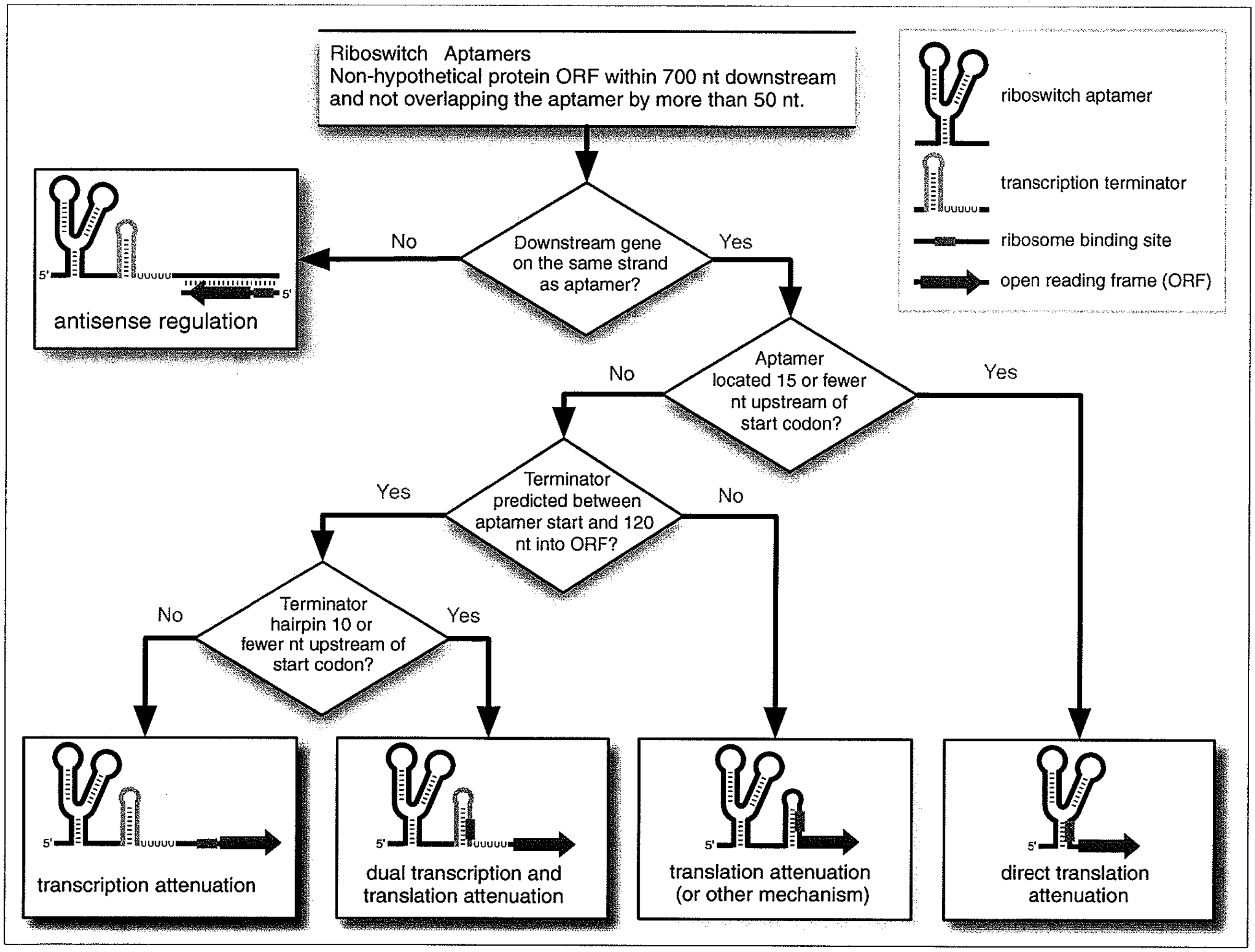

Riboswitch Aptamers
Non-hypothetical protein ORF within 700 nt downstream and not overlapping the aptamer by more than 50 nt.
riboswitch aptamer
transcription terminator
UUUUU
ribosome binding site
open reading frame (ORF)
Downstream gene on the same strand as aptamer?
No
Yes
5'
5'
antisense regulation
Aptamer located 15 or fewer nt upstream of start codon?
No
Yes
Terminator predicted between aptamer start and 120 nt into ORF?
Yes
No
Terminator hairpin 10 or fewer nt upstream of start codon?
No
Yes
transcription attenuation
dual transcription and translation attenuation
translation attenuation (or other mechanism)
direct translation attenuation

Figure 2

Riboswitch mechanism prediction scheme. The decision tree used to classify riboswitch mechanisms into five categories is shown. Depicted are OFF switches in their ligand-bound state where a P1 switching helix has formed. See the main text and Materials and methods for additional details.

ulate. However, some bacteria appear to be able to utilize other structures that may lack a canonical U-tail or consist of tandem hairpins to terminate transcription [57].

Mapping riboswitch mechanism predictions onto a phylogenetic tree (Figure 3) reveals that transcription attenuation dominates in Firmicutes and that translation attenuation is most common in other bacterial groups. The phylogenetic distribution of SAM-II riboswitch mechanisms is an exception. It is the only riboswitch aptamer that appears to be most often associated with regulatory transcription terminators in α- and β-Proteobacteria, although the mechanisms by which SAM-II aptamers control gene expression have not yet been experimentally established [18]. Transcription attenuation mechanisms may also be generally overrepresented in Fusobacteria, δ/ε-Proteobacteria, Thermatogae, and Chloroflexi species, although smaller sample sizes make these conclusions less certain.

Mechanisms that rely on sequestering the RBS within the conserved aptamer core are most common for the TPP, preQ$_1$, and SAM-I riboswitches. In the first two cases, purine-rich conserved regions near the 3' ends of the riboswitch substitute for RBS sequences. In SAM-I riboswitches, the RBS is incorporated into the 3' side of the P1 stem. Other riboswitch classes also have purine-rich conserved regions near their 3' ends with consensus sequences close to ribosome binding sites. It is not clear why direct regulation of translation attenuation is not more common in these other classes. Perhaps access to the RBS-like sequences in these aptamers is not modulated by ligand binding. Riboswitch regulation by direct translation attenuation appears to be most frequent in

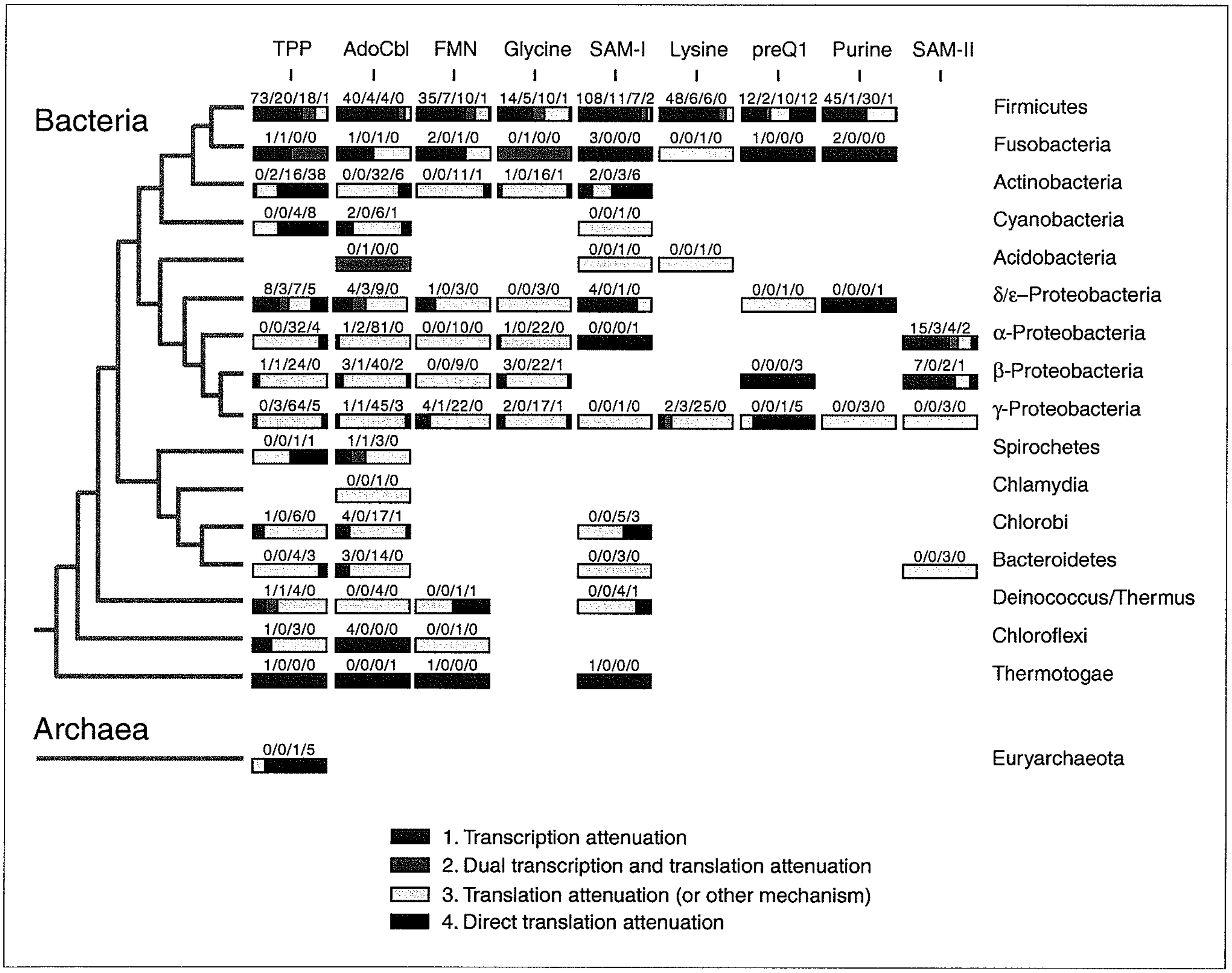


Figure 3
Riboswitch mechanisms. The mechanisms that riboswitches from different taxonomic groups use to regulate gene expression were classified on the basis of expression platform features (Figure 2). The fractions of riboswitch expression platforms in each category are displayed visually as shaded bars with the actual numbers observed written above in the order given in the legend. The phylogenetic tree on the left is described in the legend to Figure 1.

Actinobacteria and Cyanobacteria, except for the preQ$_1$ riboswitch where this mechanism is unusually prevalent, even in Firmicutes and Proteobacteria.

There do not appear to be any additional examples of riboswitches positioned for antisense regulation in this data set. An antisense arrangement may be rare because it inverts the gene control logic of the riboswitch and requires the evolutionary maintenance of a second promoter. A handful of high-scoring hits were found that appear to be functional aptamers even though they are not located upstream of genes related to the cognate metabolite. It is possible that these riboswitches affect their target genes by regulating the production or function of *trans*-acting antisense RNAs or that they have been recently orphaned by genomic rearrangements and are now pseudo-regulatory sequences.

Evaluating structure models

Constructing an RNA secondary structure model using phylogenetic sequence data requires identifying possible base-paired stems and adjusting a sequence alignment to determine whether each proposed stem appears reasonable for all representatives. This recursive refinement process has been used to create detailed comparative models of many functional RNA structures that accurately reflect later genetic, biochemical and biophysical data. However, the presence of stretches of unvarying nucleotides within an RNA structure, the tolerance of stems to some non-canonical base pairs or mismatches, and the non-negligible frequency of sequencing errors in biological databases can introduce enough uncertainty that multiple structures may seem to agree with a sequence alignment and incorrect base-paired elements may be proposed. This problem is compounded if the multiple sequence alignment is incomplete and does not yet capture all of the variation that truly exists at each nucleotide position.

Inconsistencies and ambiguities in some riboswitch aptamer models motivated us to evaluate the statistical support for base pairs in their proposed structures. We chose to use mutual information (MI) scores [58] to mathematically formalize the interdependence between sequence alignment columns that is indicative of base interactions. MI is a normalized version of covariance that represents the amount of information (in bits) gained about what base occurs at a given position from knowing the identity of a base at another position. The prediction of RNA secondary structures and tertiary interactions from covariation in sequence alignments has a long history, and the nuances of calculating and interpreting MI scores have been comprehensively covered elsewhere [59,60].

Fundamentally, columns of interacting bases must be correctly aligned and there must be variation within each column (that is, it cannot be completely conserved) in order to detect mutual information. Even when these preconditions are met, there are two difficulties with directly comparing MI scores to determine which columns in a sequence alignment truly covary. First, sequence conservation derived from the shared evolutionary histories of sequence subsets in an alignment may result in a high residual background MI score between many columns whether or not they are functionally linked. Second, alignments with fewer sequences will have more column pairs with elevated MI scores simply by chance. Simulations addressing the expected magnitudes of these two sources of error in different data sets have been explored recently in the context of protein sequence alignments [61].

In order to better gauge whether MI scores support proposed base interactions in an RNA alignment, we developed a procedure for empirically estimating their statistical significance (Figure 4). First, a phylogenetic tree is inferred from the observed RNA sequence alignment according to a model that assumes independent evolution at each position and allows for varying per-column mutation rates. Then, resampled alignments with the same topology, branch lengths, and evolutionary rates are generated. MI scores between columns in these test alignments reflect the null hypothesis that there is no covariation between positions. They implicitly correct for the evolutionary history and sample size of the real sequence alignment. Therefore, the p value significance for an observed MI score in the real alignment is the fraction of *test* alignments with higher MI scores between these two columns.

Riboswitch structures

The consensus secondary structure models of the ten riboswitch classes (Figure 5) have been updated to reflect information from newly identified aptamer variants. The purine, TPP, SAM-I, and GlcN6P riboswitch consensus structures have been drawn in accordance with their molecular structures (references in Table 1). Other riboswitch structures have been revised to be consistent with the new predictions of structure motifs and base-base interactions explained below. In all cases, previous numbering schemes for the paired helical elements (designated P1, P2, P3, and so on, beginning at the 5' end of each the aptamer) have been maintained, even when these stems do not occur in a majority of the sequences in the updated alignment. Newly discovered paired elements that do not appear in most examples of a riboswitch aptamer have not been assigned numbers.

The results of the mutual information analysis are shown superimposed on the consensus riboswitch structures. Most base-paired helices are supported by at least one contiguous base pair with a highly significant MI ($p < 0.001$), and almost all contain a base pair with at least a marginal MI significance ($p < 0.01$). No significant MI scores are present within the P2.1 and P2.2 stems observed in the crystal structures of the GlcN6P-dependent ribozyme [28,30]. However, most of the predicted base pairs in the P2.1 and P2.2 helices are between highly conserved bases that may not vary enough to produce significant covariation with their pairing partners. The MI analysis also does not support an alternative P1.1 pseudoknot

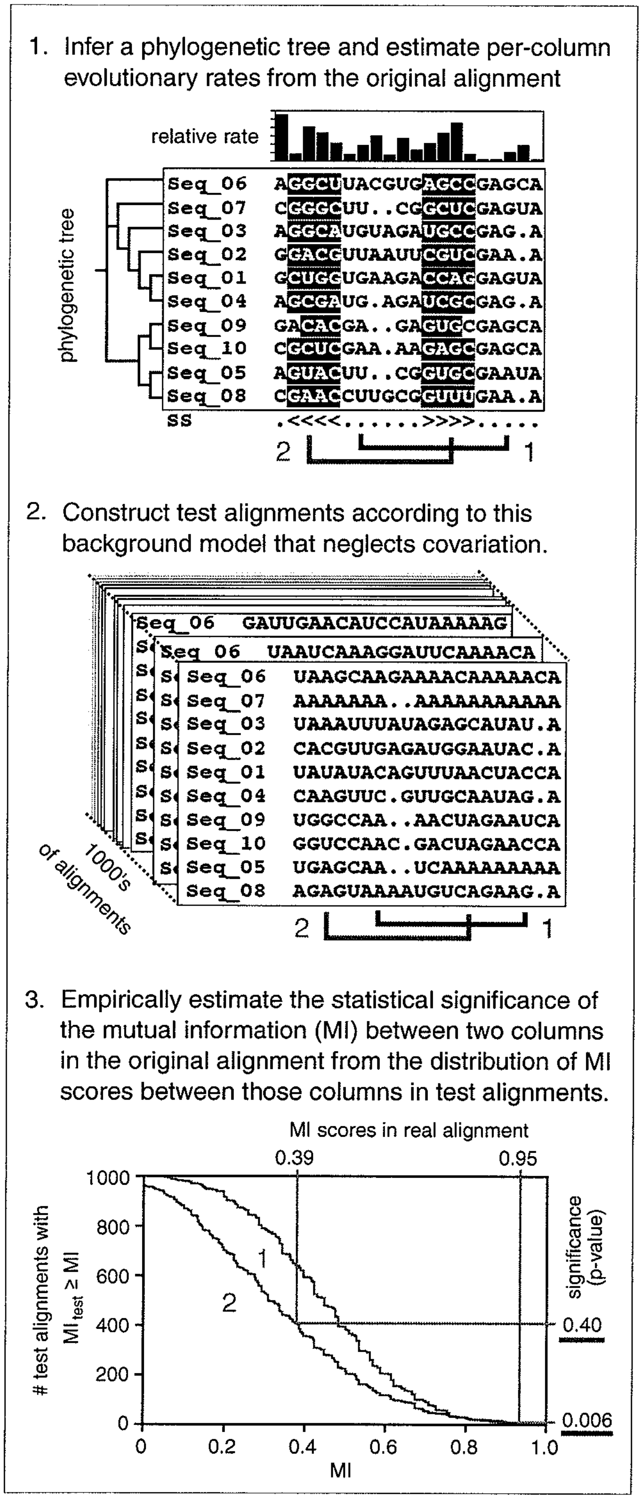

1. Infer a phylogenetic tree and estimate per-column evolutionary rates from the original alignment
relative rate
phylogenetic tree
Seq_06 AGGCUUACGUGAGCCGAGCA
Seq_07 CGGGCUU..CGGCUCGAGUA
Seq_03 AGGCAUGUAGAUGCCGAG.A
Seq_02 GGACGUUAAUUCGUCGAA.A
Seq_01 GCUGGUGAAGAGCAGGAGUA
Seq_04 AGCGAUG.AGAUCGCGAG.A
Seq_09 GACACGA..GAGUGCGAGCA
Seq_10 CGCUCGAA.AAGAGCGAGCA
Seq_05 AGUACUU..CCGUGCGAAUA
Seq_08 CGAACCUUGCCGUUUGAA.A
SS .<<<<......>>>>.....
2
1
2. Construct test alignments according to this background model that neglects covariation.
Seq_06 GAUUGAACAUCCAUAAAAAG
Seq_06 UAAUCAAAGGAUUCAAAACA
Seq_06 UAAGCAAGAAAACAAAAACA
Seq_07 AAAAAAA..AAAAAAAAAAA
Seq_03 UAAAUUUAUAGAGCAUAU.A
Seq_02 CACGUUGAGAUGGAAUAC.A
Seq_01 UAUAUACAGUUUAACUACCA
Seq_04 CAAGUUC.GUUGCAAUAG.A
Seq_09 UGGCCAA..AACUAGAAUCA
Seq_10 GGUCCAAC.GACUAGAACCA
Seq_05 UGAGCAA..UCAAAAAAAAA
Seq_08 AGAGUAAAAUGUCAGAAG.A
1000's of alignments
2
1
3. Empirically estimate the statistical significance of the mutual information (MI) between two columns in the original alignment from the distribution of MI scores between those columns in test alignments.
MI scores in real alignment
0.39
0.95
# test alignments with $MI_{test} \geq MI$
1000
800
600
400
200
0
0
0.2
0.4
0.6
0.8
1.0
MI
1
2
significance (p-value)
0.40
0.006

Figure 4

Procedure for estimating MI significance between alignment columns. See the main text and Materials and methods for a complete description of the procedure used to estimate the statistical significance of MI scores between columns in a multiple sequence alignment in order to evaluate riboswitch secondary structures and predict new base-base interactions.

(not shown) proposed on the basis of biochemical experiments where the register of the regions involved in making the P2.1 pairing is slightly shifted [29,62,63].

MI significance scores do resolve a conflict between two pairing models that have been proposed for the highly conserved B12 box of the AdoCbl riboswitch (Figure 6). One model posits that a 'facultative stem loop' forms by pairing nucleotides within the B12 box [20]. The other model proposes long-range pairings between portions of the B12 box and nucleotides more distant in RNA sequence [39]. There is only a single, marginally significant MI score that supports the formation of the 'facultative stem loop', even though this region was correctly aligned to optimally discover such interactions. The MI analysis strongly supports several base pairs in the alternative proposed structure wherein portions of the conserved B12 box form the 3' sides of the short P3 and P6 helical stems.

RNA structure motifs

Several riboswitches contain common RNA structure motifs that are recognizable from their consensus features. A GNRA tetraloop [64] that favors a pyrimidine at its second position caps P4a of most GlcN6P ribozymes. A K-turn [65,66] between P2 and P2a is conserved in SAM-I riboswitch aptamers [66]. The asymmetric bulge between helices P2a and P2b in the lysine riboswitch also fits a K-turn consensus in most sequences [67], but a number of variants appear to lack this motif. A sarcin-ricin motif [68] (a specific type of loop E motif) in the asymmetric bulge between the P2 and P2a helices of the lysine riboswitch is more highly conserved [37,67].

We also find examples of other RNA structure motifs that have not previously been reported in these riboswitch classes. The consensus features of the three terminal loops capping P2, P3, and P5 in the FMN riboswitch and the P4 loop and P6-P7 bulge in the AdoCbl riboswitch are remarkably similar. Each has two closing G-C base pairs with a strand bias, a possible U-A pair separated from the helical stem by two bulged nucleotides on the 3' side, and a terminal GNR triloop sequence that is sometimes interrupted at a specific position by an intervening base-paired helix. These characteristics strongly suggest that they adopt T-loop structures (named for the T-loop of tRNA) where the U-A forms a key *trans* Watson-Crick/Hoogsteen pair [69].

Sequence conservation in the UNR loop that closes the P5 stem in the TPP aptamer suggests that it forms a conserved U-turn [70]. As expected, there is a sharp reversal of backbone direction following this uridine, subsequent bases stack on the 3' side of the loop, and the uracil base can hydrogen bond with the phosphate group 3' of the third U-turn nucleotide in the X-ray crystal structures of *E. coli* [71,72] and *Arabidopsis thaliana* [73] riboswitches. Also, in the TPP aptamer, the conserved UGAGA sequence 3' of the P3 helix fits the UGNRA consensus for a type R1 lonepair triloop [74]. The crystal structures confirm that this motif is present with the characteristic *trans* Watson-Crick/Hoogsteen U-A closing pair around the triloop. Commonly, a tertiary interaction between the triloop G base and an outside A leads to a composite GNRA tetraloop structure. However, in this case, the pyrimidine ring from the TPP ligand intercalates into the triloop at an equivalent position.

New base-base interaction predictions

In addition to supporting almost all of the helical elements in the riboswitch structure models, the MI analysis predicts eleven additional base-pairing interactions (Figures 5 and 7). Significant MI scores between two alignment columns should be interpreted with caution. They represent a statistical correlation and do not necessarily imply hydrogen bonding between nucleobases. Correlations between adjacent nucleotides that probably represent favored base stacking patterns in helices and column pairs with many gaps where MI scores can be dominated by the presence and absence of nucleotides rather than their base identities have been ignored. It is also possible to observe high mutual information between two bases that do not interact if several separate structure motifs with their own specific sequence requirements can substitute for each other in a functional RNA, as is seen for GNRA, UNCG, and CUUG tetraloops in 16S rRNA [59].

Furthermore, the estimates of MI significance rely on a phylogenetic tree reconstruction method that may not adequately capture the evolution of these RNA sequences, especially for the shorter riboswitch alignments. Even assuming that the estimated *p* values are completely accurate, there are 4,950 possible combinations of columns in an alignment with 100 columns, and that would imply that, on average, 5 pairs with a MI significance of ≤0.01 will be observed by chance. Some columns that are known to be base paired do not have MI scores this significant. In light of this noisy background, we manually screened MI predictions and concentrated on interacting columns that seem to have structural relevance.

The identities of interacting bases in a functional RNA are constrained during evolution. They can mutate only to other base pairs that preserve the local geometry of the sugar-phosphate backbone and any hydrogen bonds that are important for maintaining structure and function. Generally, only one of the three planar edges of a nucleobase participates in any given interaction: the Watson-Crick face (WC), Hoogsteen face (H), or sugar edge (SE). A systematic study of RNA structures has produced isostericity matrices [75] that tabulate which of the possible 16 base pairs should be interchangeable (in terms of C1'-C1' distances) when two nucleobases are interacting between different combinations of these three base edges and when the glycosidic bonds on both sides of the pair are *cis* or *trans* with respect to each other. The pairs of bases conserved at some of the new correlated positions in riboswitches suggest unusual non-Watson-Crick interactions, and this isostericity framework can be used to tentatively assign possible geometries to the newly predicted base pairs (Figure 7).

In the TPP riboswitch, there is significant MI between the two bases directly 5' of P3 and 3' of P3a that could bridge this helical junction. This correlation was highly significant (p = 0.0002) in an alignment of all TPP riboswitch sequences. However, re-examination of the alignment showed that the predominant A-G and U-A pairs mainly occurred in the 552 sequences that have the optional P3a stem-loop. In fact, there is no correlation between these columns in the remaining 355 sequences that lack P3a. Exchange of U-A and A-G pairs is most consistent with a *cis* H/WC edge interaction between these two bases. These pairs are also isosteric in a *trans* H/H geometry, but this configuration involves only a single hydrogen bond, and there are four other isosteric nucleobase combinations that are not observed. Both pair geometries imply that either the sugar-phosphate backbones of the interacting bases are in a parallel orientation or that they are anti-parallel, with one of the bases adopting a rare *syn* glycosidic bond rotation. It may be necessary for these bases to assume an unusual geometry to accommodate the P3a helix at this location.

The molecular resolution structures of TPP riboswitches do not impinge on this prediction, as each of these constructs lacks P3a [71-73]. On the basis of the consensus structure, it is possible to further predict that when the P3a helix is present it will coaxially stack on the P2 helix as part of a type C three-way helical junction [76] wherein P3a, P2, and P3 are assigned P1, P2, and P3 roles, respectively. The molecular structures show a diagnostic feature of this configuration even in the absence of P3a: the J13 motif sequence (corresponding to the conserved UGAGA) forms a pseudohairpin that makes adenine base contacts to the minor groove of the motif's P1 helix (P2 of the riboswitch). Furthermore, there is space in the crystal structure to accommodate P3a cohelically stacking on P2, and this would place P3a parallel to and offset from P3, as is expected for this common three-way junction geometry.

Figure 5 *(see following page)*
Riboswitch aptamer structures. The consensus secondary structure models based on expanded riboswitch sequence alignments are depicted according to the symbols defined in the inset. Each structure is further annotated with RNA structure motifs and the statistical significances (*p* values) of the mutual information scores between base-paired alignment columns. New predictions of interacting bases from the MI analysis are numbered and indicated by asterisks. More detailed descriptions of these predictions are provided in Figure 7.

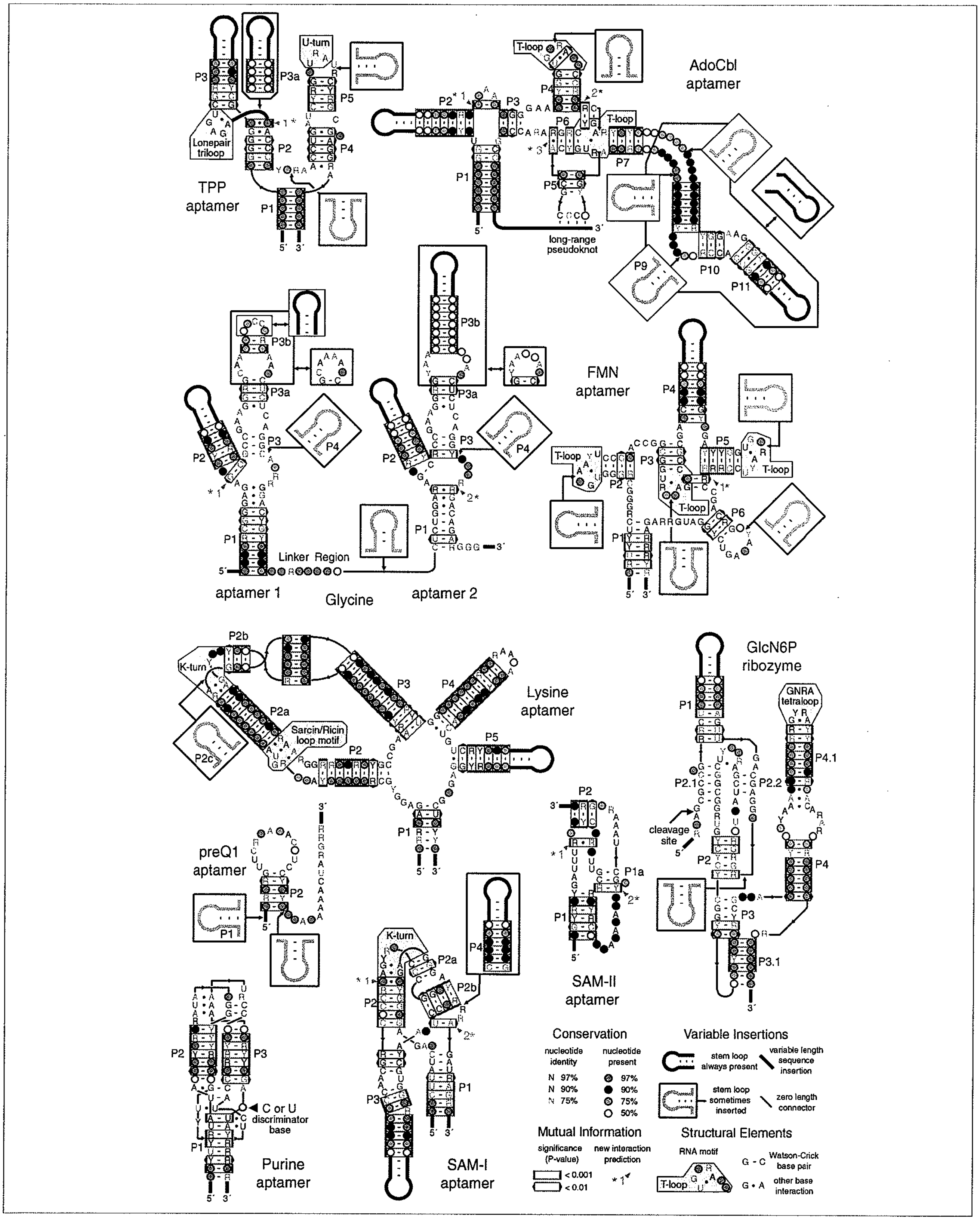


Figure 5 *(see legend on previous page)*

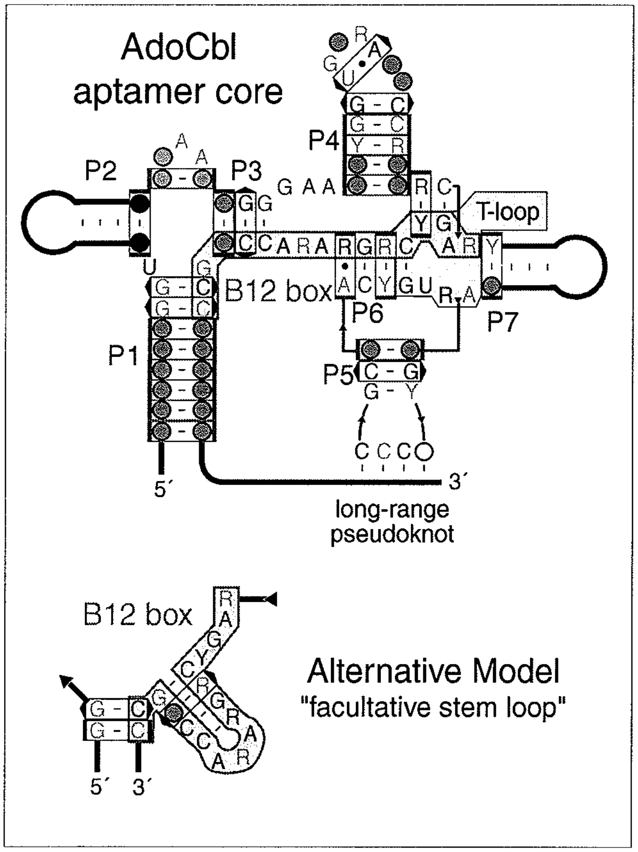


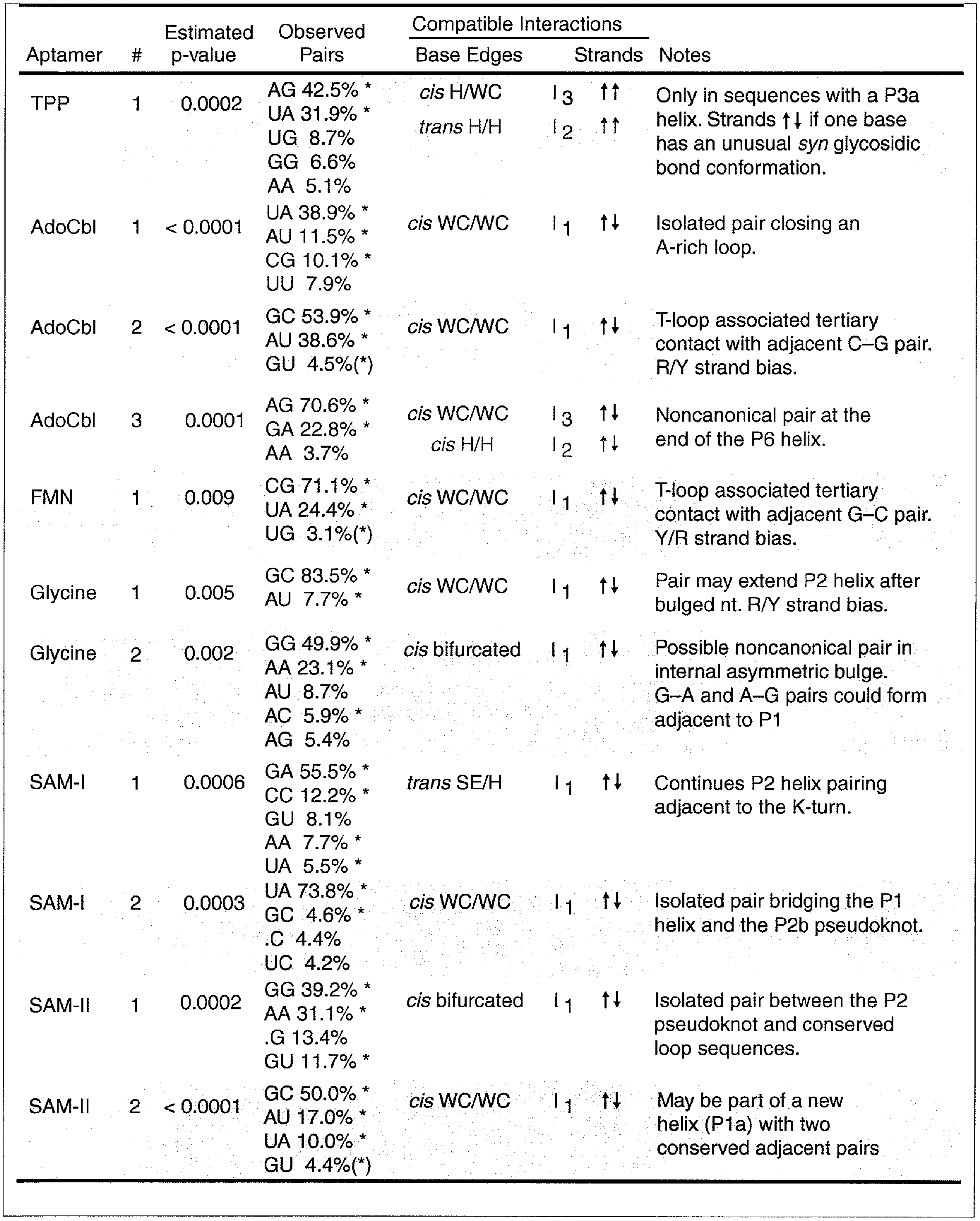
Figure 6
Comparison of B12 box structure models. In addition to the model of the AdoCbl riboswitch aptamer structure presented here [39], an alternative model that folds the highly conserved B12 box sequence (highlighted in red) into a 'facultative stem-loop' has been proposed [20]. The core of the AdoCbl riboswitch aptamer is shown with abbreviated peripheral helices and without the optional P8-P10-P11 domain for comparison with the alternative secondary structure model. The upper model is supported by multiple base pairs with significant MI scores between B12 box bases and remote positions. In it, a portion of the B12 box also forms part of an internal T-loop motif between P6 and P7. Each diagram uses the symbols described in the legend to Figure 5.

Three new base interactions are predicted in AdoCbl riboswitch aptamers. A lone WC base pair ($p < 0.0001$) seems to enclose the conserved A-rich sequence between the P2 and P3 helices. A highly significant MI score ($p < 0.0001$) also supports a WC pair with purine/pyrimidine strand bias between the nucleotide directly 3' of the P4 helix and a position within the two nucleotide 3' bulge of the P6-P7 T-loop motif. The adjacent nucleotides in this strand and the T-loop bulge could form a highly conserved, cohelical C-G base pair. Similar long-range Watson-Crick base-pairing interactions to these two bulged nucleotides are common with 'type-II' T-loops [69]. The final new prediction in the AdoCbl riboswitch is a non-canonical G-A or A-G pair ($p = 0.0001$) that probably assumes a *cis* WC/WC geometry to continue base stacking with the P6 helix. These pairs are also isosteric in a *cis* H/H geometry, but this geometry seems less likely to be conserved because it involves only a single hydrogen bond.

The FMN riboswitch may contain a strikingly similar T-loop interaction. The nucleotide directly 3' of its P5 helix can form a Watson-Crick pair ($p = 0.009$) with a pyrimidine/purine strand bias to the 3' bulge of the T-loop motif that caps P3. An adjacent G-C base pair is also possible here between highly conserved nucleotides in the strand and T-loop bulge. In both the AdoCbl and FMN riboswitches, the stem-loops adjacent to this predicted interaction have exactly five paired nucleotides and are capped by a second T-loop motif. Although the second T-loop does not seem to be directly relevant to this predicted pairing interaction, the double T-loop substructure that these riboswitches have in common suggests that significant similarity exists between their overall tertiary folds even though they recognize very different ligand molecules.

The MI analysis suggests two new base-base interactions in the glycine riboswitch. The first is a WC pair ($p = 0.005$) with purine/pyrimidine strand bias at the base of the P2 stem of the first aptamer. If this pair cohelically stacks with the P2 stem, then it would often require a bulged nucleotide on the 5' side of the composite helix. The second interaction is a predicted G-G or A-A homopurine pair ($p = 0.002$) that might adopt a *cis* bifurcated geometry within the central bulge of the second aptamer. Bifurcated pairs hydrogen bond between an exocyclic functional group on one base and the edge of the other base, and they are consequently intermediate between two edge geometries (possibly *cis* WC/WC and *trans* WC/H in this case). If this pair forms, it suggests that the two bases on each strand between it and the P1 stem may form G-A and A-G pairs. Both of these putative interactions are maintained in the opposite aptamer of the glycine riboswitch. However, the nucleotides at the corresponding positions are less variable, which may explain why they were not detected a second time by the MI analysis.

Two new base-pairing contacts are predicted for SAM-I riboswitches. The first occurs at the end of the P2 helix adjacent to the conserved G-A and A-G pairs of the K-turn motif. This pair has a highly significant MI score ($p = 0.0006$) and mainly varies from G-A to C-C, which is most compatible with a *trans* SE/H base interaction within this cohelical stacking context. Noncanonical pairs with this configuration are known to occur frequently adjacent to K-turns in other functional RNA structures [77]. The second predicted interaction ($p = 0.0003$) is an unexpected long-range *cis* WC/WC base pair between the base directly upstream of the 5' side of the P2b pseudoknot and the base directly upstream of the P1 3' strand. After originally discovering these new interactions from sequence analysis, we were able to verify that both interactions occur with the predicted configurations in the X-ray crystal structure of a minimized version of the *Thermoanaerobacter tengcongensis metF* SAM-I riboswitch [78].



Aptamer	#	Estimated p-value	Observed Pairs	Compatible Interactions		Notes
				Base Edges	Strands	
TPP	1	0.0002	AG 42.5% * UA 31.9% * UG 8.7% GG 6.6% AA 5.1%	*cis* H/WC *trans* H/H	I_3 ↑↑ I_2 ↑↑	Only in sequences with a P3a helix. Strands ↑↓ if one base has an unusual *syn* glycosidic bond conformation.
AdoCbl	1	< 0.0001	UA 38.9% * AU 11.5% * CG 10.1% * UU 7.9%	*cis* WC/WC	I_1 ↑↓	Isolated pair closing an A-rich loop.
AdoCbl	2	< 0.0001	GC 53.9% * AU 38.6% * GU 4.5%(*)	*cis* WC/WC	I_1 ↑↓	T-loop associated tertiary contact with adjacent C–G pair. R/Y strand bias.
AdoCbl	3	0.0001	AG 70.6% * GA 22.8% * AA 3.7%	*cis* WC/WC *cis* H/H	I_3 ↑↓ I_2 ↑↓	Noncanonical pair at the end of the P6 helix.
FMN	1	0.009	CG 71.1% * UA 24.4% * UG 3.1%(*)	*cis* WC/WC	I_1 ↑↓	T-loop associated tertiary contact with adjacent G–C pair. Y/R strand bias.
Glycine	1	0.005	GC 83.5% * AU 7.7% *	*cis* WC/WC	I_1 ↑↓	Pair may extend P2 helix after bulged nt. R/Y strand bias.
Glycine	2	0.002	GG 49.9% * AA 23.1% * AU 8.7% AC 5.9% * AG 5.4%	*cis* bifurcated	I_1 ↑↓	Possible noncanonical pair in internal asymmetric bulge. G–A and A–G pairs could form adjacent to P1
SAM-I	1	0.0006	GA 55.5% * CC 12.2% * GU 8.1% AA 7.7% * UA 5.5% *	*trans* SE/H	I_1 ↑↓	Continues P2 helix pairing adjacent to the K-turn.
SAM-I	2	0.0003	UA 73.8% * GC 4.6% * .C 4.4% UC 4.2%	*cis* WC/WC	I_1 ↑↓	Isolated pair bridging the P1 helix and the P2b pseudoknot.
SAM-II	1	0.0002	GG 39.2% * AA 31.1% * .G 13.4% GU 11.7% *	*cis* bifurcated	I_1 ↑↓	Isolated pair between the P2 pseudoknot and conserved loop sequences.
SAM-II	2	< 0.0001	GC 50.0% * AU 17.0% * UA 10.0% * GU 4.4%(*)	*cis* WC/WC	I_1 ↑↓	May be part of a new helix (P1a) with two conserved adjacent pairs

Figure 7 *(see legend on next page)*

Figure 7 *(see previous page)*
New base-base interaction predictions. For each numbered and asterisked prediction in Figure 5 the statistical significance (*p* value) of the mutual information between the two alignment columns is shown, followed by the relative frequencies with which specific combinations of bases are observed in those columns. Base pair geometries and isostericity groups compatible with the asterisked pairs are described in more detail elsewhere [75]. These descriptions include the relative orientations of the glycosidic bonds across the pair (*cis* or *trans*), the edges of each base that interact (WC, Watson-Crick; H, Hoogsteen; SE, sugar edge; bifurcated, intermediate between two edges), and the relative backbone strand geometry (parallel or anti-parallel) assuming both glycosidic bonds are in default *anti* conformations.

The MI analysis predicts two new base-base interactions in the SAM-II riboswitch. A homopurine G-G or A-A pair (p = 0.0002) could form between two positions in the bulge between P1 and the 5' strand of the P2 pseudoknot. This pair may adopt a *cis* bifurcated geometry. A Watson-Crick base pair ($p < 0.0001$) may also exist between the last nucleotide in the central loop that is contained within the P1 stem and a downstream position. This pair could be extended into a short helical element (P1a) if the adjacent, conserved C-G and G-C base pairs also form canonical WC pairs and an intervening base is bulged out.

Conclusion

The ten metabolite-sensing riboswitch classes surveyed here are widespread and versatile gene control elements. The conserved secondary structure models of these riboswitch aptamers have been revised to include information from additional sequence variants. These models incorporate newly recognized RNA structure motifs, including a double T-loop substructure that is conserved in AdoCbl and FMN aptamers, and specify new sites where the insertion of unconserved RNA domains is possible. Furthermore, an analysis of mutual information scores using an evolutionarily informed background model has enabled the prediction of new base-base interactions in several riboswitch aptamers. These refinements should improve the accuracy of future computational searches for riboswitches as the automated annotation of functional RNAs in genomic sequences becomes more routine [19]. They will also inform and validate ongoing efforts to determine the molecular resolution structures of riboswitch aptamers.

It is believed that some metabolite-binding riboswitch classes may be descended from the RNA World [79] and that others may be more recent evolutionary innovations [80], but the exact provenance of each riboswitch class is unclear. Significant uncertainty also remains about what physiological and evolutionary forces affect riboswitch use by modern organisms. Particularly, there are unexplained differences in the distributions and preferred regulatory mechanisms of riboswitches across contemporary bacteria. Riboswitches found in Firmicutes (low G+C Gram-positive bacteria) predominantly regulate transcription attenuation, whereas translation attenuation mechanisms are most prevalent in other groups. Overall, riboswitches also appear to be more common in Firmicutes than other bacterial groups.

One of the more interesting aspects of the riboswitch phylogenetic profile is that it outlines gaps and holes in the known distributions of riboswitch classes. Some of these apparently vacant regulatory niches may be occupied by regulatory proteins that fulfill the same role or by extreme structural variants of these riboswitch classes that are not detectable with current RNA homology search techniques. Other gaps could harbor new aptamer classes that recognize the same metabolite as a known riboswitch class. The discovery of SAM-II riboswitches in α-Proteobacteria [18], which are almost devoid of SAM-I riboswitches, sets a precedent for this latter scenario. The existence of a third SAM riboswitch in some lactic acid bacteria species [81], a subdivision of the Firmicutes, suggests that new riboswitch classes may occupy empty regulatory niches that exist at an even finer taxonomic resolution.

Materials and methods

Computational analysis

In-house Perl scripts were used to organize the execution of other software tools, compute various statistics, and maintain local relational databases of genome and gene information. Many of these scripts rely on Bioperl [82], and the Bio::Graphics module was particularly useful for visualizing the genomic contexts of riboswitch matches.

Riboswitch identification

Covariance models were trained on sequence alignments adapted from various sources (Table 1) using the Infernal software package (version 0.55) [83]. Heuristic filtering techniques [16] were used to accelerate CM searches of microbial sequences in the RefSeq database (version 12) [84] and environmental shotgun sequences from an acid mine drainage community [85], the Sargasso Sea [25], and Minnesota soil and whale fall sites [86]. CM searches for TPP riboswitches were also conducted against the plant and fungal portions of the RefSeq database (version 13).

The regulatory potentials of putative riboswitch aptamers were assessed by examining their genomic contexts. To uniformly predict gene functions, protein domains were assigned to COGs (orthologous gene clusters) [87] using RPS-BLAST and scoring matrices from the Conserved Domain Database (CDD) [88]. The plausibility of putative aptamer structures was assessed by computationally aligning hits to the original CM with Infernal and manually examining divergent RNA structures. Using these two complementary criteria, we established trusted CM score cutoffs. All hits in the microbial

RefSeq database above these thresholds were judged to be functional riboswitches. Since gene context information is not available for most environmental sequences, hits from these data sets were included only if they had CM scores above the trusted threshold. Additional low-scoring sequences from the RefSeq database were also included when their genomic contexts and alignments strongly indicated that they were functional riboswitches.

To verify that this approach efficiently recovers known riboswitches, the final results were compared to a list of TPP riboswitches compiled in a comparative genomics analysis of thiamin metabolic genes and this regulatory RNA element [48]. The new searches successfully found all TPP riboswitches that had been previously identified in the set of complete microbial genomes analyzed in both studies. They also discovered a small number of TPP riboswitches upstream of thiamin-related genes (for example, a *pnuC* homolog in *Helicobacter pylori* and *thiM* in *Lactococcus lactis*) in genomes examined by the former study that had not yet been reported.

For the glycine riboswitch, a single aptamer covariance model and a tandem model containing both the first and second aptamers were used to separately identify matches. Every aptamer that is part of a tandem configuration was found by the single aptamer CM search, and cases of lone aptamers were noted. For consensus structure and MI calculations only the tandem glycine aptamer alignment was considered, but the complete set of lone and tandem aptamer glycine riboswitches were included in the expression platform analysis. Expression platform counts for other riboswitch classes that rarely occur in tandem were not corrected.

Mechanism classification

Expression platforms were classified according to the scheme in Figure 2 for a subset of the riboswitch matches found in complete and unfinished microbial genomes. Aptamer sequences with more than 95% pairwise identity at reference columns (positions where ≥50% of the weighted sequences in the alignment do not contain a gap) were omitted to avoid biasing statistics with duplicate sequences. Riboswitches with suspect gene annotations where >60 nucleotides (nt) of an open reading frame (ORF) on the same strand overlapped the aptamer or >700 nt separated the aptamer and the nearest downstream ORF were also screened out. Most of these cases appear to result from incorrect start codon choices, overpredictions of hypothetical ORFs, or missing annotation of real genes. The remaining sequences constituted the expression platform data set, and sequences beginning at the 5' end of each aptamer and continuing through the first 120 nt of the downstream ORF were extracted for further analysis.

Riboswitches where the downstream gene was on the opposite strand were examined as candidates for antisense regulation. Other riboswitches were classified as directly regulating translation initiation when the downstream gene's start codon was within 15 nt of the end of the conserved aptamer core structure (usually the P1 paired element). The remaining expression platforms were scanned with the local RNA secondary structure prediction program Rnall (version 1.1) [89] for intrinsic transcription terminators with a scanning window of 50 nt, a U-tail weight threshold of 4.0, a U-tail pairing stability cutoff of -8.3 kcal/mol, and default settings for other parameters. Riboswitches with a terminator predicted in their expression platform sequence were assigned transcription attenuation mechanisms. These riboswitches were classified as also regulating translation if the distance between the terminator hairpin and the gene's start codon is no more than 10 nt. Expression platforms that did not match any of the above criteria are assumed to employ translation attenuation mechanisms.

Rnall and distance parameters were calibrated by comparing expression platform predictions to expert predictions for a large and phylogenetically diverse collection of TPP riboswitches [48]. Rnall correctly predicts 46 out of 52 terminators in this data set with only 3 predictions of terminators in sequences not manually evaluated as containing a terminator (a sensitivity of 88% and an accuracy of 94%). The three false positives resemble terminators and may be functional, whereas the terminators that Rnall misses usually have large hairpins with poor thermodynamic stabilities. Overall, the decision tree classifies 159 out of 180 TPP riboswitch expression platforms (88%) correctly into the category assigned in the control set.

Consensus secondary structures

We manually adjusted the covariance model alignments of riboswitch aptamers while refining their consensus secondary structures. In particular, bases taking part in pseudoknotted pairings that cannot be represented by CMs were shifted to accurately represent these interactions. Bases flanking gapped consensus columns, which are sometimes ambiguously spread out across many possible positions by the alignment algorithm, were also systematically condensed into a minimum number of overall consensus columns. As new structure motifs and base-base interactions became evident, the alignments were adjusted to reflect these new constraints. Riboswitch sequences in the final alignments were weighted using Infernal's internal implementation of the GSC algorithm [90] to reduce biases from duplicate and similar sequences before calculating consensus structure statistics.

Mutual information significance

Duplicate sequences were purged and columns with >50% gaps were removed from riboswitch alignments prior to the MI analysis, and, if necessary, alignments were further pruned to the 300 most diverse sequences (as judged by pairwise base differences). A customized version of the program Rate4Site (version 2.01) [91] with modified output options was used to simultaneously estimate distances and per-column rates of evolution according to a gamma distributed model with at least 16 rate categories and a phylogenetic tree created with Jukes-Cantor distances that treated gaps as missing information. The resulting trees, rates, and distances were used to simulate 10,000 resampled alignments starting from an arbitrary ancestral sequence. Then, gaps and sequence weights were re-inserted into each of these derivative alignments at the same positions that they occupied in the original alignment.

Mutual information was calculated between column pairs for all alignments according to standard formulas [60], taking into account sequence weights and treating gaps as a fifth character state. The resampled alignments were used to estimate what the MI score distribution would have been if the bases present in each column had evolved independently, without covariation constraints. The p value significance of the actual MI between two columns is the fraction of the resampled alignments that have a greater MI score than the value observed between those two columns in the real alignment.

Abbreviations

AdoCbl, adenosylcobalamin; CM, covariance model; FMN, flavin mononucleotide; GlcN6P, glucosamine-6-phosphate; H, Hoogsteen face; MI, mutual information; nt, nucleotides; ORF, open reading frame; preQ$_1$, 7-aminoethyl 7-deazaguanine; RBS, ribosome binding site; SAM, *S*-adenosylmethionine; SE, sugar edge; TPP, thiamin pyrophosphate; UTR, untranslated region; WC, Watson-Crick face.

Authors' contributions

JEB designed the computational analyses, carried out the comparative studies, and created the figures. JEB and RRB interpreted the results and wrote the manuscript.

Additional data files

The following additional data files are available with the online version of this article. Additional data file 1 contains sequence alignments of the riboswitch aptamer data sets annotated with new base-base interactions in Stockholm format. Additional data file 2 contains sequence alignments of the riboswitch aptamer data sets annotated with new base-base interactions in HTML format.

Acknowledgements

We thank BC Grove for early work on the programs used for mutual information calculations, GM Emilsson for assistance drafting Figure 2, and other members of the Breaker lab for helpful discussions. We also thank E Westhof for detailed feedback on an early draft, X-F Wan for porting RNALL to MacOS X at our request, and I Mayrose and T Pupko for answering questions regarding Rate4Site. JEB was supported by a Howard Hughes Medical Institute Predoctoral Fellowship. RRB is supported by the Howard Hughes Medical Institute and this project was also funded by NIH grants GM068819 and DK070270.

References

1. Mandal M, Breaker RR: **Gene regulation by riboswitches.** *Nat Rev Mol Cell Biol* 2004, **5**:451-463.
2. Winkler WC, Breaker RR: **Regulation of bacterial gene expression by riboswitches.** *Annu Rev Microbiol* 2005, **59**:487-517.
3. Winkler WC: **Metabolic monitoring by bacterial mRNAs.** *Arch Microbiol* 2005, **183**:151-159.
4. Tucker BJ, Breaker RR: **Riboswitches as versatile gene control elements.** *Curr Opin Struct Biol* 2005, **15**:342-348.
5. Winkler WC, Nahvi A, Roth A, Collins JA, Breaker RR: **Control of gene expression by a natural metabolite-responsive ribozyme.** *Nature* 2004, **428**:281-286.
6. Mandal M, Lee M, Barrick JE, Weinberg Z, Emilsson GM, Ruzzo WL, Breaker RR: **A glycine-dependent riboswitch that uses cooperative binding to control gene expression.** *Science* 2004, **306**:275-279.
7. Welz R, Breaker RR: **Ligand binding and gene control characteristics of tandem riboswitches in *Bacillus anthracis*.** *RNA* 2007, **13**:573-582.
8. Richter-Dahlfors AA, Andersson DI: **Cobalamin (vitamin B$_{12}$) repression of the Cob operon in *Salmonella typhimurium* requires sequences within the leader and the first translated open reading frame.** *Mol Microbiol* 1992, **6**:743-749.
9. Grundy FJ, Henkin TM: **The S box regulon: a new global transcription termination control system for methionine and cysteine biosynthesis genes in Gram-positive bacteria.** *Mol Microbiol* 1998, **30**:737-749.
10. Miranda-Ríos J, Morera C, Taboada H, Dávalos A, Encarnacíon S, Mora J, Soberón M: **Expression of thiamin biosynthetic genes (*thiCOGE*) and production of symbiotic terminal oxidase *cbb*$_3$ in *Rhizobium etli*.** *J Bacteriol* 1997, **179**:6887-6893.
11. Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, Sampath R: **RNAMotif, an RNA secondary structure definition and search algorithm.** *Nucleic Acids Res* 2001, **29**:4724-4735.
12. Abreu-Goodger C, Merino E: **RibEx: a web server for locating riboswitches and other conserved bacterial regulatory elements.** *Nucleic Acids Res* 2005, **33**:W690-W692.
13. Eddy SR, Durbin R: **RNA sequence analysis using covariance models.** *Nucleic Acids Res* 1994, **22**:2079-2088.
14. Weinberg Z, Ruzzo WL: **Faster genome annotation of non-coding RNA families without loss of accuracy.** In *Proceedings of the Eighth Annual International Conference on Computational Molecular Biology (RECOMB). San Diego, CA. March 27-31, 2004* Edited by: Philip E. Bourne ACM Press, New York, NY; 2004:243-251.
15. Weinberg Z, Ruzzo WL: **Exploiting conserved structure for faster annotation of non-coding RNAs without loss of accuracy.** *Bioinformatics* 2004, **20**:i334-i341.
16. Weinberg Z, Ruzzo WL: **Sequence-based heuristics for faster annotation of non-coding RNA families.** *Bioinformatics* 2006, **22**:35-39.
17. Barrick JE, Sudarsan N, Weinberg Z, Ruzzo WL, Breaker RR: **6S RNA is a widespread regulator of eubacterial RNA polymerase that resembles an open promoter.** *RNA* 2005, **11**:774-784.
18. Corbino KA, Barrick JE, Lim J, Welz R, Tucker BJ, Puskarz I, Mandal M, Rudnick ND, Breaker RR: **Evidence for a second class of *S*-adenosylmethionine riboswitches and other regulatory RNA motifs in alpha-proteobacteria.** *Genome Biol* 2005, **6**:R70.
19. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A: **Rfam: annotating non-coding RNAs in complete genomes.** *Nucleic Acids Res* 2005, **33**:D121-D124.
20. Vitreschak AG, Rodionov DA, Mironov AA, Gelfand MS: **Regulation of the vitamin B$_{12}$ metabolism and transport in bacteria by a conserved RNA structural element.** *RNA* 2003, **9**:1084-1097.
21. Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS: **Regulation of lysine biosynthesis and transport genes in bacteria: yet another RNA riboswitch?** *Nucleic Acids Res* 2003, **31**:6748-6757.
22. Mandal M, Boese B, Barrick JE, Winkler WC, Breaker RR: **Riboswitches control fundamental biochemical pathways in *Bacillus subtilis* and other bacteria.** *Cell* 2003, **113**:577-586.
23. Barrick JE, Corbino KA, Winkler WC, Nahvi A, Mandal M, Collins J, Lee M, Roth A, Sudarsan N, Jona I, *et al.*: **New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control.** *Proc Natl Acad Sci USA* 2004, **101**:6421-6426.
24. Weinberg Z, Barrick JE, Yao Z, Roth A, Kim JN, Gore J, Wang JX, Lee ER, Block KF, Sudarsan N, *et al.*: **Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline.** *Nucleic Acids Res* 2007, **35**:4809-4819.

25. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu DY, Paulsen I, Nelson KE, Nelson W, *et al.*: **Environmental genome shotgun sequencing of the Sargasso Sea.** *Science* 2004, **304:**66-74.
26. DeLong EF: **Microbial community genomics in the ocean.** *Nat Rev Microbiol* 2005, **3:**459-469.
27. Hampel KJ, Tinsley MM: **Evidence for preorganization of the *glmS* ribozyme ligand binding pocket.** *Biochemistry* 2006, **45:**7861-7871.
28. Klein DJ, Ferre-D'Amare AR: **Structural basis of *glmS* ribozyme activation by glucosamine-6-phosphate.** *Science* 2006, **313:**1752-1756.
29. McCarthy TJ, Plog MA, Floy SA, Jansen JA, Soukup JK, Soukup GA: **Ligand requirements for *glmS* ribozyme self-cleavage.** *Chem Biol* 2005, **12:**1221-1226.
30. Cochrane JC, Lipchock SV, Strobel SA: **Structural investigation of the GlmS ribozyme bound to its catalytic cofactor.** *Chem Biol* 2007, **14:**97-105.
31. Rieder R, Lang K, Graber D, Micura R: **Ligand-induced folding of the adenosine deaminase A-riboswitch and implications on riboswitch translational control.** *Chembiochem* 2007, **8:**896-902.
32. Gilbert SD, Stoddard CD, Wise SJ, Batey RT: **Thermodynamic and kinetic characterization of ligand binding to the purine riboswitch aptamer domain.** *J Mol Biol* 2006, **359:**754-768.
33. Wickiser JK, Winkler WC, Breaker RR, Crothers DM: **The speed of RNA transcription and metabolite binding kinetics operate an FMN riboswitch.** *Mol Cell* 2005, **18:**49-60.
34. Wickiser JK, Cheah MT, Breaker RR, Crothers DM: **The kinetics of ligand binding by an adenine-sensing riboswitch.** *Biochemistry* 2005, **44:**13404-13414.
35. Lemay JF, Penedo JC, Tremblay R, Lilley DM, Lafontaine DA: **Folding of the adenine riboswitch.** *Chem Biol* 2006, **13:**857-868.
36. Winkler W, Nahvi A, Breaker RR: **Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression.** *Nature* 2002, **419:**952-956.
37. Sudarsan N, Wickiser JK, Nakamura S, Ebert MS, Breaker RR: **An mRNA structure in bacteria that controls gene expression by binding lysine.** *Genes Dev* 2003, **17:**2688-2697.
38. Nahvi A, Sudarsan N, Ebert MS, Zou X, Brown KL, Breaker RR: **Genetic control by a metabolite binding mRNA.** *Chem Biol* 2002, **9:**1043-1049.
39. Nahvi A, Barrick JE, Breaker RR: **Coenzyme B_{12} riboswitches are widespread genetic control elements in prokaryotes.** *Nucleic Acids Res* 2004, **32:**143-150.
40. Roth A, Winkler WC, Regulski EE, Lee BW, Lim J, Jona I, Barrick JE, Ritwik A, Kim JN, Welz R, *et al.*: **A riboswitch selective for the queuosine precursor preQ$_1$ contains an unusually small aptamer domain.** *Nat Struct Mol Biol* 2007, **14:**308-317.
41. Sudarsan N, Barrick JE, Breaker RR: **Metabolite-binding RNA domains are present in the genes of eukaryotes.** *RNA* 2003, **9:**644-647.
42. Kubodera T, Watanabe M, Yoshiuchi K, Yamashita N, Nishimura A, Nakai S, Gomi K, Hanamoto H: **Thiamine-regulated gene expression of *Aspergillus oryzae thiA* requires splicing of the intron containing a riboswitch-like domain in the 5'-UTR.** *FEBS Lett* 2003, **555:**516-520.
43. Cheah MT, Wachter A, Sudarsan N, Breaker RR: **Control of alternative RNA splicing and gene expression by eukaryotic riboswitches.** *Nature* 2007, **447:**497-500.
44. Yarnell WS, Roberts JW: **Mechanism of intrinsic transcription termination and antitermination.** *Science* 1999, **284:**611-615.
45. Gusarov I, Nudler E: **The mechanism of intrinsic transcription termination.** *Mol Cell* 1999, **3:**495-504.
46. Mandal M, Breaker RR: **Adenine riboswitches and gene activation by disruption of a transcription terminator.** *Nat Struct Mol Biol* 2004, **11:**29-35.
47. Merino E, Yanofsky C: **Transcription attenuation: a highly conserved regulatory strategy used by bacteria.** *Trends Genet* 2005, **21:**260-264.
48. Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS: **Comparative genomics of thiamin biosynthesis in procaryotes: new genes and regulatory mechanisms.** *J Biol Chem* 2002, **277:**48949-48959.
49. Vitreschak AG, Rodionov DA, Mironov AA, Gelfand MS: **Riboswitches: the oldest mechanism for the regulation of gene expression?** *Trends Genet* 2004, **20:**44-50.
50. Nou XW, Kadner RJ: **Adenosylcobalamin inhibits ribosome binding to *btuB* RNA.** *Proc Natl Acad Sci USA* 2000, **97:**7190-7195.
51. Altman S, Wesolowski D, Guerrier-Takada C, Li Y: **RNase P cleaves transient structures in some riboswitches.** *Proc Natl Acad Sci USA* 2005, **102:**11284-11289.
52. Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS: **Comparative genomics of the methionine metabolism in Gram-positive bacteria: a variety of regulatory systems.** *Nucleic Acids Res* 2004, **32:**3340-3353.
53. Gottesman S: **Stealth regulation: biological circuits with small RNA switches.** *Genes Dev* 2002, **16:**2829-2842.
54. Rodionov DA, Dubchak I, Arkin A, Alm E, Gelfand MS: **Reconstruction of regulatory and metabolic pathways in metal-reducing delta-proteobacteria.** *Genome Biol* 2004, **5:**R90.
55. Sudarsan N, Hammond MC, Block KF, Welz R, Barrick JE, Roth A, Breaker RR: **Tandem riboswitch architectures exhibit complex gene control functions.** *Science* 2006, **314:**300-304.
56. Vitreschak AG, Rodionov DA, Mironov AA, Gelfand MS: **Regulation of riboflavin biosynthesis and transport genes in bacteria by transcriptional and translational attenuation.** *Nucleic Acids Res* 2002, **30:**3141-3151.
57. Unniraman S, Prakash R, Nagaraja V: **Conserved economics of transcription termination in eubacteria.** *Nucleic Acids Res* 2002, **30:**675-684.
58. Chiu DK, Kolodziejczak T: **Inferring consensus structure from nucleic acid sequences.** *Comput Appl Biosci* 1991, **7:**347-352.
59. Gutell RR, Power A, Hertz GZ, Putz EJ, Stormo GD: **Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods.** *Nucleic Acids Res* 1992, **20:**5785-5795.
60. Lindgreen S, Gardner PP, Krogh A: **Measuring covariation in RNA alignments: physical realism improves information measures.** *Bioinformatics* 2006, **22:**2988-2995.
61. Gloor GB, Martin LC, Wahl LM, Dunn SD: **Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions.** *Biochemistry* 2005, 44:7156-7165.
62. Soukup GA: **Core requirements for *glmS* ribozyme self-cleavage reveal a putative pseudoknot structure.** *Nucleic Acids Res* 2006, **34:**968-975.
63. Jansen JA, McCarthy TJ, Soukup GA, Soukup JK: **Backbone and nucleobase contacts to glucosamine-6-phosphate in the *glmS* ribozyme.** *Nat Struct Mol Biol* 2006, **13:**517-523.
64. Heus HA, Pardi A: **Structural features that give rise to the unusual stability of RNA hairpins containing GNRA loops.** *Science* 1991, **253:**191-194.
65. Klein DJ, Schmeing TM, Moore PB, Steitz TA: **The kink-turn: a new RNA secondary structure motif.** *EMBO J* 2001, **20:**4214-4221.
66. Winkler WC, Grundy FJ, Murphy BA, Henkin TM: **The GA motif: an RNA element common to bacterial antitermination systems, rRNA, and eukaryotic RNAs.** *RNA* 2001, **7:**1165-1172.
67. Grundy FJ, Lehman SC, Henkin TM: **The L box regulon: Lysine sensing by leader RNAs of bacterial lysine biosynthesis genes.** *Proc Natl Acad Sci USA* 2003, **100:**12057-12062.
68. Leontis NB, Stombaugh J, Westhof E: **Motif prediction in ribosomal RNAs: lessons and prospects for automated motif prediction in homologous RNA molecules.** *Biochimie* 2002, **84:**961-973.
69. Nagaswamy U, Fox GE: **Frequent occurrence of the T-loop RNA folding motif in ribosomal RNAs.** *RNA* 2002, **8:**1112-1119.
70. Gutell RR, Cannone JJ, Konings D, Gautheret D: **Predicting U-turns in ribosomal RNA with comparative sequence analysis.** *J Mol Biol* 2000, **300:**791-803.
71. Serganov A, Polonskaia A, Phan AT, Breaker RR, Patel DJ: **Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch.** *Nature* 2006, **441:**1167-1171.
72. Edwards TE, Ferre-D'Amare AR: **Crystal structures of the *Thi*-box riboswitch bound to thiamine pyrophosphate analogs reveal adaptive RNA-small molecule recognition.** *Structure* 2006, **14:**1459-1468.
73. Thore S, Leibundgut M, Ban N: **Structure of the eukaryotic thiamine pyrophosphate riboswitch with its regulatory ligand.** *Science* 2006, **312:**1208-1211.
74. Lee JC, Cannone JJ, Gutell RR: **The lonepair triloop: a new motif in RNA structure.** *J Mol Biol* 2003, **325:**65-83.
75. Leontis NB, Stombaugh J, Westhof E: **The non-Watson-Crick base pairs and their associated isostericity matrices.** *Nucleic Acids Res* 2002, **30:**3497-3531.
76. Lescoute A, Westhof E: **Topology of three-way junctions in folded RNAs.** *RNA* 2006, **12:**83-93.
77. Lescoute A, Leontis NB, Massire C, Westhof E: **Recurrent struc-**

tural RNA motifs, isostericity matrices and sequence alignments. *Nucleic Acids Res* 2005, **33:**2395-2409.

78. Montange RK, Batey RT: **Structure of the *S*-adenosylmethionine riboswitch regulatory mRNA element.** *Nature* 2006, **441:**1172-1175.
79. Breaker RR: **Riboswitches and the RNA World.** In *The RNA World* 3rd edition. Edited by: Gesteland RF, Cech TR, Atkins JF. Woodbury, NY: Cold Spring Harbor Laboratory Press; 2005:89-108.
80. Gold L, Brody E, Heilig J, Singer B: **One, two, infinity: Genomes filled with aptamers.** *Chem Biol* 2002, **9:**1259-1264.
81. Fuchs RT, Grundy FJ, Henkin TM: **The S_{MK} box is a new SAM-binding RNA for translational regulation of SAM synthetase.** *Nat Struct Mol Biol* 2006, **13:**226-233.
82. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, *et al.*: **The Bioperl toolkit: Perl modules for the life sciences.** *Genome Res* 2002, **12:**1611-1618.
83. **Infernal: Inference of RNA Alignments** [http://infernal.janelia.org/]
84. Pruitt KD, Tatusova T, Maglott DR: **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2005, **33:**D501-504.
85. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF: **Community structure and metabolism through reconstruction of microbial genomes from the environment.** *Nature* 2004, **428:**37-43.
86. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, *et al.*: **Comparative metagenomics of microbial communities.** *Science* 2005, **308:**554-557.
87. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, *et al.*: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4:**41.
88. Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z, *et al.*: **CDD: a Conserved Domain Database for protein classification.** *Nucleic Acids Res* 2005, **33:**D192-196.
89. Wan XF, Xu D: **Intrinsic terminator prediction and its application in *Synechococcus* sp. WH8102.** *J Comput Sci Tech* 2005, **20:**465-482.
90. Gerstein M, Sonnhammer ELL, Chothia C: **Volume changes in protein evolution.** *J Mol Biol* 1994, **236:**1067-1078.
91. Mayrose I, Graur D, Ben-Tal N, Pupko T: **Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior.** *Mol Biol Evol* 2004, **21:**1781-1791.
92. Madigan MT, Martinko JM, Parker J: *Brock Biology of Microorganisms* 10th edition. Upper Saddle River, NJ: Pearson Education, Inc; 2003.
93. Mira A, Pushker R, Legault BA, Moreira D, Rodriguez-Valera F: **Evolutionary relationships of *Fusobacterium nucleatum* based on phylogenetic analysis and comparative genomics.** *BMC Evol Biol* 2004, **4:**50.
94. Winkler WC, Nahvi A, Sudarsan N, Barrick JE, Breaker RR: **An mRNA structure that controls gene expression by binding *S*-adenosylmethionine.** *Nat Struct Biol* 2003, **10:**701-707.
95. Batey RT, Gilbert SD, Montange RK: **Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine.** *Nature* 2004, **432:**411-415.
96. Serganov A, Yuan YR, Pikovskaya O, Polonskaia A, Malinina L, Phan AT, Hobartner C, Micura R, Breaker RR, Patel DJ: **Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs.** *Chem Biol* 2004, **11:**1729-1741.
97. Noeske J, Richter C, Grundl MA, Nasiri HR, Schwalbe H, Wohnert J: **An intermolecular base triple as the basis of ligand specificity and affinity in the guanine- and adenine-sensing riboswitch RNAs.** *Proc Natl Acad Sci USA* 2005, **102:**1372-1377.

nature
biotechnology

Riboswitches as antibacterial drug targets

Kenneth F Blount[1] & Ronald R Breaker[1–3]

New validated cellular targets are needed to reinvigorate antibacterial drug discovery. This need could potentially be filled by riboswitches—messenger RNA (mRNA) structures that regulate gene expression in bacteria. Riboswitches are unique among RNAs that serve as drug targets in that they have evolved to form structured and highly selective receptors for small drug-like metabolites. In most cases, metabolite binding to the receptor represses the expression of the gene(s) encoded by the mRNA. If a new metabolite analog were designed that binds to the receptor, the gene(s) regulated by that riboswitch could be repressed, with a potentially lethal effect to the bacteria. Recent work suggests that certain antibacterial compounds discovered decades ago function at least in part by targeting riboswitches. Herein we will summarize the experiments validating riboswitches as drug targets, describe the existing technology for riboswitch drug discovery and discuss the challenges that may face riboswitch drug discoverers.

There is increasing concern that the current antibacterial drug arsenal includes primarily decades-old chemical scaffolds that target a very narrow spectrum of cellular processes. It has been highlighted recently that antibiotics drug development has produced only one new chemical scaffold in the past 30 years, and currently prescribed antibiotics collectively disrupt the function of only four bacterial life processes[1]. Perhaps as a consequence, antibiotic resistance is now emerging at an alarmingly rapid pace, with indications that even the most recently approved antibiotics could soon be ineffective[2]. Sustained success in the long-term battle against bacterial pathogens will require the identification of new chemical scaffolds that target other cellular processes. One potentially vulnerable process is the regulation of gene expression by metabolite-sensing RNAs called riboswitches.

In many bacteria, the expression of a number of genes crucial to metabolite biosynthesis or transport is regulated by mRNA structures called riboswitches[3–5]. Typically found in the 5′-untranslated region (5′-UTR) of certain bacterial mRNAs, members of each known riboswitch class form a structured receptor, or 'aptamer', that has evolved to bind to a specific fundamental metabolite. If the cognate metabolite is not present when the 5′-UTR is transcribed ("GENE ON," **Fig. 1a,b**), the riboswitch in most cases folds into a structure that does not interfere with the expression of the adjacent open reading frame (ORF). When present at a sufficiently high concentration, the metabolite binds to the riboswitch receptor, which induces the formation of a structure in the nascent mRNA that represses the expression of the ORF. This structure can be a terminator hairpin, which halts RNA synthesis before the ORF can be synthesized ("GENE OFF," **Fig. 1a**) or a hairpin that sequesters the Shine-Dalgarno sequence and prevents the ribosome from binding to the mRNA and translating the ORF ("GENE OFF," **Fig. 1b**). Because the gene or group of genes regulated by a riboswitch is usually involved in the synthesis or transport of its cognate metabolite, riboswitches are direct regulators of cellular metabolite concentrations.

Researchers have reported 12 different classes of riboswitches, and members of each class bind to the same metabolite and share a highly conserved sequence and secondary structure. For example, the 5′-UTR of the *tenA* operon in *Staphylococcus aureus* contains a 177-nucleotide RNA element that is homologous to a class of riboswitches that selectively bind the coenzyme thiamine pyrophosphate (TPP, **Fig. 1c**)[6,7]. Similar TPP riboswitches have been identified in the genomes of species from most bacterial phyla[7]. Phylogenetic sequence comparison suggests that, when bound to TPP, the minimal receptor domain of all TPP riboswitches is a secondary structure comprised of five helices (P1 through P5) joined by three stretches of primarily unpaired nucleotides (J2-3, J2-4 and J4-5, **Fig. 1c,d**)[7]. Structural probing data with representative TPP riboswitches are consistent with this structure[6,8,9]. Strikingly, the identities of the nucleotides in J2-3, J2-4 and J4-5 are almost universally conserved among TPP riboswitches, as are some of the nucleotides in the helical regions (**Fig. 1d**). A similar level of sequence and secondary structure conservation is observed for other riboswitch classes, which implies that all riboswitches of a given class form a common fold to recognize their cognate ligand (**Fig. 2a**).

In some bacterial pathogens, a riboswitch regulates the expression of a gene that is essential for survival or virulence (**Fig. 2b**)[10]. It is debatable whether a positive therapeutic outcome could be achieved by targeting riboswitches that control only virulence genes. However, it seems reasonable to predict that a metabolite mimic designed to target a riboswitch controlling genes essential for survival should have a lethal effect, provided that the mimic cannot functionally replace the natural metabolite. For instance, in *S. aureus* the *metK* gene, which encodes *S*-adenosylmethionine (SAM) synthetase, is predicted to be essential for both survival[11] and virulence[12], and its expression is regulated by a class 1 SAM-binding riboswitch located in its 5′-UTR[13–15]. A metabolite mimic that binds to that riboswitch could, in principle, inhibit *S. aureus* growth and/or

[1]Department of Molecular, Cellular and Developmental Biology, [2]Department of Molecular Biophysics and Biochemistry and [3]Howard Hughes Medical Institute, Yale University, 266 Whitney Avenue, New Haven, Connecticut 06520, USA. Correspondence should be addressed to K.F.B. (ken.blount@yale.edu) or R.R.B. (ronald.breaker@yale.edu).

Published online 11 December 2006; doi:10.1038/nbt1268

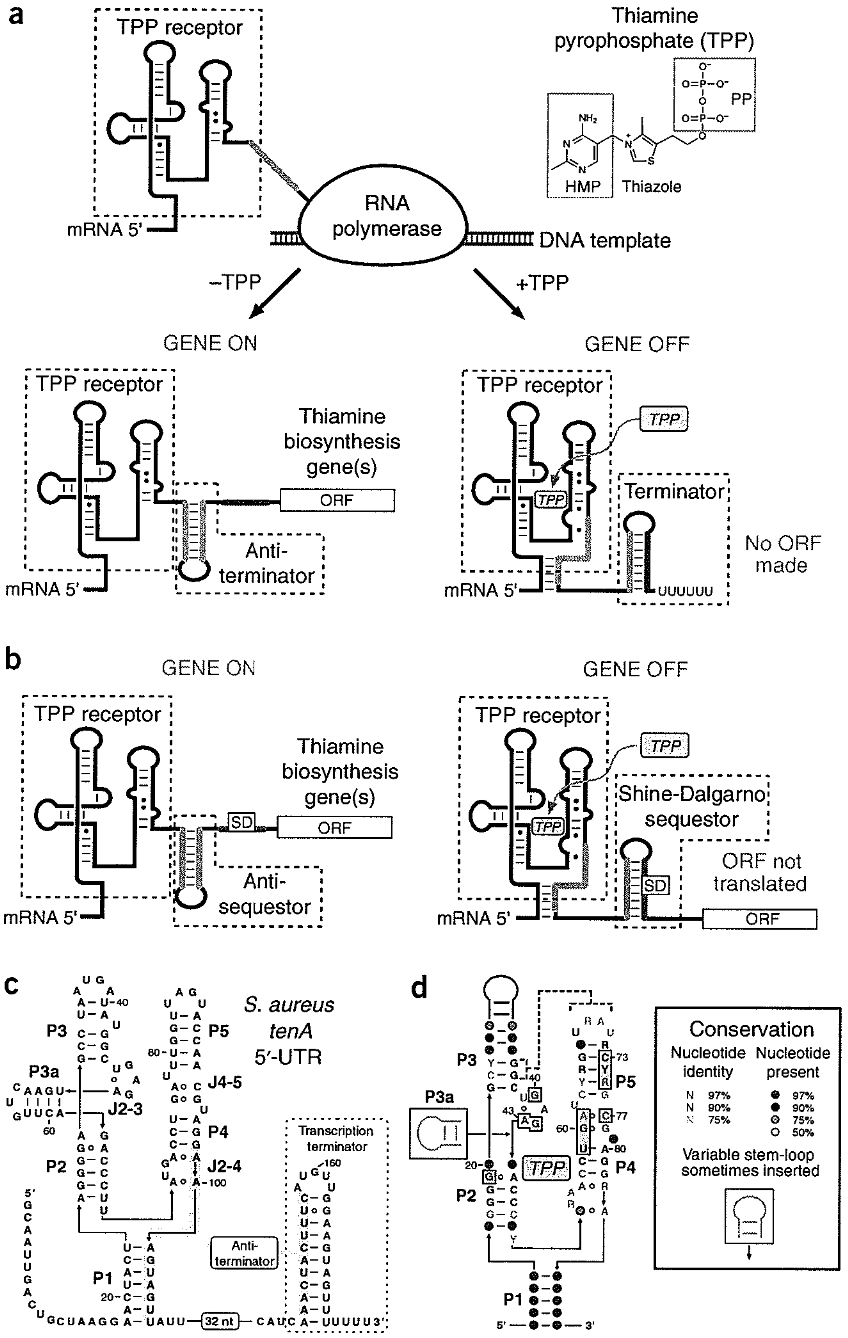


Figure 1 TPP-binding riboswitches regulate gene expression. (**a**) Transcription termination mechanism used by some TPP riboswitches to regulate gene expression. If TPP is not present during the synthesis of the 5′-UTR (GENE ON), an antiterminator hairpin (blue) forms which does not interfere with gene expression. When present at higher concentrations (GENE OFF), TPP binds to the receptor and stabilizes helix 1 (P1, see **c**), thereby allowing a terminator hairpin to form. (**b**) A translation inhibition mechanism for riboswitch gene regulation. If TPP is not present (GENE ON), an antisequestor hairpin (blue) forms that does not interfere with gene expression. At higher concentrations (GENE OFF), TPP binds to the receptor and stabilizes P1, thereby allowing a hairpin to form that sequesters the Shine-Dalgarno sequence (SD) and prevents the ribosome from translating the ORF. (**c**) The sequence and secondary structure model for the repressed-state TPP riboswitch from the *S. aureus tenA* 5′-UTR. The sequence is numbered according to the predicted start site of transcription. An additional 21 nucleotides (not depicted) reside between nucleotide 177 and the *tenA* start codon. The putative antiterminator hairpin that forms in the absence of ligand is shaded blue. (**d**) A consensus sequence and secondary structure model for the ligand-bound form of all bacterial TPP riboswitches. The consensus sequence and structural model were determined by phylogenetic sequence comparison (Breaker, R.R. *et al.*, unpublished data) and X-ray structural models for representatives of this riboswitch class[25,26]. P3a is not present in some sequences. Nucleotides that form direct or metal ion–mediated contacts to TPP are boxed in yellow. Dashed brackets indicate an interhelical tertiary interaction that is observed in the 3D structure (**Fig. 3b**).

virulence by repressing *metK* and by preventing SAM biosynthesis. In other words, the compound would be an agonistic drug that causes gene repression, even when the cell is starved for SAM.

Even when a riboswitch does not regulate the expression of a known essential gene, it could still be a drug target if it regulates several genes whose collective repression would be deleterious. For example, in *Listeria monocytogenes* a riboswitch that binds coenzyme B_{12} (adenosylcobalamin, AdoCbl) regulates the expression of the entire *cobB* operon, which contains 19 genes whose collective repression would prevent biosynthesis and transport of AdoCbl[16]. Presumably, a metabolite mimic that targets this riboswitch could induce AdoCbl starvation and halt bacterial growth. Additionally, some bacterial genomes carry multiple riboswitches of the same class, each regulating a different operon. *Bacillus anthracis* has 17 class 1 SAM-binding riboswitches that collectively regulate the expression of 36 genes responsible for the majority of sulfur metabolism. Since SAM riboswitches are highly conserved, a single compound conceivably could repress all 36 genes, halt sulfur metabolism and thereby prevent growth.

There are many examples where genetic evidence suggests that riboswitches would be excellent targets for antibacterial drug discovery. Some classes of riboswitches, such as those that sense TPP or flavin mononucleotide (FMN), are found in many different bacterial species and might be targets for broad-spectrum antibacterial drugs, whereas other riboswitch classes are more sparsely distributed and might be targets for selective drugs. In the following sections, we will highlight the current status of riboswitch target validation for antibacterial drug development. As well, we will survey the existing technology for riboswitch drug discovery and speculate on the challenges that will need to be overcome to develop riboswitch-targeting antibiotics.

Riboswitches as distinctive RNA drug targets

RNA is already recognized as a clinically useful antibacterial drug target. Many currently marketed antibacterial drugs, including the two most recently approved[17], target ribosomal RNA (rRNA) structures[18,19].

Numerous other RNA motifs also have been explored as potential drug targets[20]. In most of these examples, association of the ligand with the RNA target is a fortuitous interaction, rather than the natural function of the RNA. As a consequence, the design of new ligands directed at these targets can be problematic due to a lack of selectivity[21].

Riboswitches are fundamentally different RNA drug targets, in that they have evolved as structured receptors for the purpose of binding low-molecular weight ligands. As a consequence, riboswitches form ligand-receptor interfaces with a level of structural complexity and selectivity that approaches that of proteins. For example, the receptor of a guanine-binding riboswitch from *Bacillus subtilis* forms a three-dimensional (3D) structure in which the ligand is almost completely enveloped (**Fig. 3a**)[22–24]. Guanine intercalates between two aromatic base triads, and four additional riboswitch nucleotides form hydrogen-bonding interactions that recognize each polar functional group of guanine. The intimacy of this interaction enables the riboswitch to discriminate against even closely related purine analogs[22], and the sequence conservation of the nucleotides that form the binding pocket implies that all guanine riboswitches form the same core receptor structure (**Fig. 3a**)[22].

Members of other riboswitch classes form similarly intricate ligand-receptor interfaces and show equally high levels of ligand selectivity. For instance, TPP riboswitches form a receptor comprised of one subdomain that recognizes every polar functional group of the 4-amino-5-hydroxymethyl-2-methylpyrimidine (HMP) moiety (**Fig. 3b**, left panel) and a second subdomain that coordinates two metal ions and several water molecules to bind the negatively charged pyrophosphate (PP) moiety (**Fig. 3b**, right panel)[25,26]. These two subdomains are positioned so that the riboswitch can sense the length of its ligand. The receptor is probably not selective for the structure of the thiazole ring, since its only contact with the receptor is a long-range electrostatic interaction. In a similar example, the 3D structure and biochemical data for class 1 SAM-binding riboswitches reveal a receptor in which nearly every functional group of SAM is important for binding (**Fig. 3c**)[27,28].

Collectively, these data provide compelling evidence that riboswitches form structured receptors that are among the most selective of any RNA drug target. Thus, it is likely that riboswitch-targeting compounds could be designed that are highly selective and do not bind to other cellular targets. It is also anticipated that the 3D structure models for riboswitch receptors will enable the rational design of such compounds.

Several known antibacterial compounds function by targeting riboswitches. Historical validation exists that riboswitches can be antibacterial drug targets. Pyrithiamine is an analog of thiamine that inhibits the growth of several bacterial and fungal species[29,30]. Until recently, the toxicity mechanism of pyrithiamine was not well understood. Like thiamine, pyrithiamine is readily phosphorylated inside cells to pyrithiamine pyrophosphate[31] (PTPP, **Fig. 4a**), which differs from TPP only in that the central thiazole ring is replaced by a pyridinium ring. Remarkably, PTPP binds to several TPP riboswitches *in vitro* with comparable affinity to TPP and represses the expression of a reporter gene fused to a TPP riboswitch inside bacteria[9]. This suggests the possibility that, in its phosphorylated form, pyrithiamine inhibits bacterial or fungal

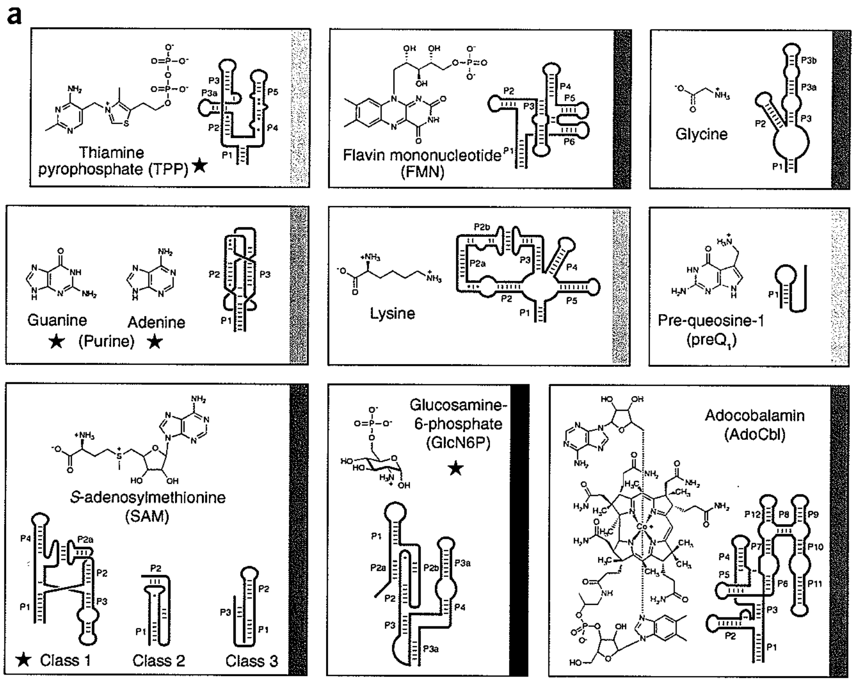


Figure 2 The distribution of known riboswitch classes in selected bacterial pathogens. (**a**) The secondary structure models of the receptor domains of 12 riboswitch classes and the chemical structures of the metabolites they bind. The expected protonation states at physiological conditions are shown for each functional group. Three different structural classes of riboswitches have been reported that recognize SAM. A star denotes a riboswitch class for which a representative 3D structure has been reported. Approximately a half-dozen additional classes of conserved RNA motifs that may be riboswitches have been identified in bacteria[61,62]. (**b**) Selected human bacterial pathogens that carry riboswitches. The number of representatives of each riboswitch class that is found in each species is given, followed in parentheses by the total number of genes regulated by those riboswitches. The red numbers indicate that at least one of the genes regulated by that riboswitch is predicted to be essential for survival or virulence. Some bacteria have two glycine aptamers (asterisks) in the same 5′-UTR that cooperatively bind two molecules of glycine to regulate the downstream operon[63].

b

Riboswitch target metabolite:	TPP	FMN	AdoCbl	Purine	SAM1	SAM2	SAM3	Lysine	GlcN6P	Glycine	PreQ1
Acinetobacter baumannii	1(3)	1(1)	1(1)	–	–	–	–	–	–	1*(1)	–
Bacillus anthracis	6(19)	2(5)	1(1)	6(9)	17(36)	–	–	4(4)	1(2)	1(1)	2(5)
Brucella melitensis	2(11)	1(1)	2(5)	–	–	1(1)	–	–	–	1*(3)	–
Enterococcus faecalis	2(5)	–	2(4)	1(2)	–	–	1(1)	1(1)	1(1)	–	2(3)
Escherichia coli	3(11)	1(1)	1(2)	–	–	–	–	1(1)	–	–	–
Francisella tularensis	1(1)	1(5)	–	–	–	1(1)	–	–	–	–	1(1)
Haemophilus influenza	3(11)	1(1)	–	–	–	–	–	1(1)	–	1*(1)	1(1)
Helicobacter pylori	1(2)	–	–	–	–	–	–	–	–	–	–
Listeria monocytogenes	2(5)	1(1)	2(20)	2(3)	7(14)	–	–	1(1)	1(1)	1(3)	1(1)
Mycobacterium tuberculosis	2(6)	–	2(4)	–	–	–	–	–	–	2*(3)	–
Pseudomonas aeruginosa	1(1)	1(2)	5(24)	–	–	–	–	–	–	–	–
Salmonella enterica	3(11)	1(1)	2(22)	–	–	–	–	–	–	–	–
Staphylococcus aureus	2(7)	2(5)	–	1(2)	4(6)	–	–	2(2)	1(1)	1(3)	2(4)
Streptococcus pneumoniae	4(11)	1(4)	–	1(2)	–	–	1(1)	–	–	1(1)	–
Vibrio cholerae	2(7)	1(1)	1(1)	–	–	–	–	3(3)	–	1(1)	–
Yersinia pestis	3(8)	1(1)	1(2)	–	–	–	–	–	–	–	–

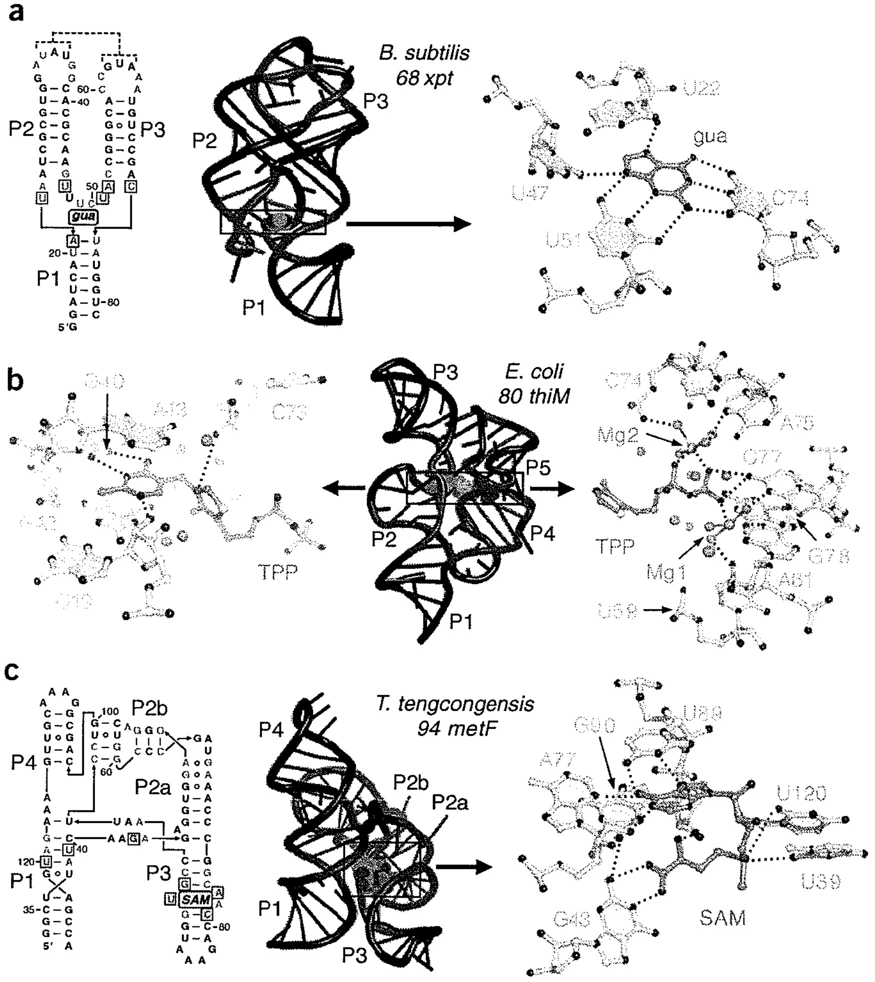


Figure 3 Structural models of ligand-bound riboswitch receptors. Nucleotides shaded in red are conserved in more than 90% of the known representatives. (**a**) The secondary structure (left) and crystal structure (center) of the repressed-state guanine riboswitch receptor from the *B. subtilis xpt* 5′-UTR. The sequence is numbered according to the predicted start site of transcription. Nucleotides that form direct contacts to guanine are boxed in yellow and dashed brackets indicate tertiary loop-loop interactions. The binding pocket (right) recognizes guanine by a series of predicted hydrogen-bonding interactions (dashed lines). (**b**) The crystal structure of the TPP-bound receptor from the *E. coli thiM* 5′-UTR (center). The binding pocket recognizes the HMP moiety (left) by a series of predicted hydrogen-bonding interactions (dashed lines) and the pyrophosphate moiety (right) is recognized through coordination to water molecules (blue) and two magnesium ions (gray). (**c**) The secondary structure (left) and crystal structure (center) of the repressed-state SAM riboswitch receptor from the *Thermoanaerobacter tengcongensis metF* 5′-UTR. The sequence is numbered according to the predicted start site of transcription. Nucleotides that form direct contacts to SAM are boxed in yellow. The binding pocket (right) recognizes SAM by a series of predicted hydrogen-bonding and electrostatic interactions (dashed lines).

growth by repressing one or more TPP riboswitch-regulated genes in these organisms.

Indeed, several strains of *B. subtilis*, *Escherichia coli* and *Aspergillus oryzae* that were cultured to resist the effect of pyrithiamine have a mutation in a conserved region of a TPP riboswitch[9]. In each case, the riboswitch that was mutated normally regulates the expression of thiamine biosynthesis genes. Moreover, the mutations disrupt ligand binding to the riboswitch *in vitro* and prevent the riboswitch from repressing a reporter gene in bacteria. Thus, it is likely that resistance to pyrithiamine is conferred through derepression of thiamine biosynthesis. These findings support the model that the TPP riboswitch is the cellular target for PTPP.

A similar case history exists for two analogs of lysine that inhibit bacterial growth. L-aminoethylcysteine[32] (AEC) and DL-4-oxalysine[33] (**Fig. 4b**) were originally reported as analogs of lysine that inhibit the growth of certain Gram-positive bacteria. More recently, it was revealed that both compounds bind to the lysine riboswitch from the *lysC* gene of *B. subtilis*, and both can repress a lysine riboswitch-regulated reporter gene in *B. subtilis*[34]. Furthermore, several strains of *B. subtilis* that were cultured to resist the effect of AEC or DL-4-oxalysine have mutations in the *lysC* riboswitch, each of which disrupts receptor formation and ligand binding *in vitro*, which prevents the riboswitch from regulating reporter gene expression in *B. subtilis*[34–37]. These studies suggest that AEC and DL-4-oxalysine inhibit bacterial growth at least in part by targeting a lysine riboswitch.

The antibacterial riboflavin analog roseoflavin (**Fig. 4c**) may also target a riboswitch. Roseoflavin inhibits the growth of several Gram-positive bacterial species through a mechanism that is reported to involve repression of riboflavin biosynthesis[38,39]. In these species, all of the genes involved in riboflavin biosynthesis are expressed in a single operon, regulated by a single FMN-binding riboswitch[40,41]. A reasonable prediction is that this riboswitch is the cellular target of roseoflavin. According to this model, roseoflavin binds to an FMN riboswitch from *B. subtilis in vitro* (Breaker, R.R. *et al.*, unpublished results), and roseoflavin-resistant strains of *B. subtilis* or *Lactococcus lactis* have mutations in the FMN riboswitch that result in overproduction of riboflavin[42–44]. Thus, like the TPP and lysine riboswitches, the FMN riboswitch may be the target of an antibacterial compound for which the mechanism of action was not previously understood. Thus, riboswitch-targeting compounds appear to have existed for decades without a clear understanding of their molecular targets.

Discovering new riboswitch-targeting antibacterials

For drug companies to maximize their chances for success in discovering new riboswitch-targeting antibacterials, they need to apply modern technologies such as rational drug design and high-throughput screening. There are already many reasons to be optimistic that riboswitch-targeting compounds could be rationally designed. Examination of the 3D structure of the ligand-bound TPP riboswitch[25,26] (**Fig. 3b**) suggests how it could accommodate other ligands like PTPP. As noted above, the central thiazole ring of TPP is not directly recognized by the receptor. Instead, it is surrounded by several water molecules. Because PTPP differs only in the identity of the central ring, it likely docks into the receptor pocket in the same way as TPP[45]. The prediction, therefore, is that other

TPP analogs that have chemically diverse functional groups linking the HMP and pyrophosphate moieties would also bind to and activate this riboswitch.

Antibacterial compounds that target other classes of riboswitches have also been identified by rational design. AEC and DL-4-oxalysine (**Fig. 4b**) are both modified at C4, relative to lysine, which suggests that other C4-modified lysine analogs could bind the riboswitch and inhibit bacterial growth. Based on this hypothesis, researchers in our laboratory have identified two additional antibacterial lysine analogs (**Fig. 4b**, inset) that inhibit the growth of *B. subtilis* most likely by targeting a lysine riboswitch[34]. In a similar study, the 3D structure of the ligand-bound guanine riboswitch was used to design several guanine analogs that bind to a guanine riboswitch from *B. subtilis* equally as well as guanine and that inhibit the growth of clinically relevant pathogens (Breaker, R.R. *et al.*, unpublished results).

The results described above suggest that, as more 3D structures of riboswitches are reported, additional avenues for rational design will become available for the discovery of riboswitch-targeting antibacterials. However, it is not clear how easily the more advanced strategies used to rationally design protein-binding compounds can be applied to riboswitches. The characteristics of some riboswitches including conformational flexibility, near total ligand encapsulation and the inherent chemical differences in RNA chemistry might cause considerable challenges for rational drug designers who seek to target metabolite-binding RNAs.

High-throughput screening has also been applied in the search for new riboswitch-activating compounds. Riboswitches that bind to glucosamine-6-phosphate (GlcN6P) are unique in that they form an RNA enzyme, or ribozyme, that self-cleaves when the cognate ligand is bound[46]. Two different reports[47,48] have described the development of a high-throughput assay to screen for other compounds that could also activate self-cleavage of this riboswitch. In one approach, ligand-induced RNA self-cleavage was detected by measuring the fluorescence polarization of a fluorescein label conjugated to the 5′-end of a GlcN6P riboswitch from *B. subtilis*[47]. A second approach used fluorescence resonance energy transfer (FRET) to detect the cleavage of a bimolecular GlcN6P riboswitch engineered from the natural riboswitch found in *S. aureus*[48]. In both cases, the assays were performed in microplates using automated pipetting methods that can easily be adapted to screen large libraries of compounds.

In principle, a similar high-throughput screen could be designed for riboswitches that do not self-cleave. One strategy would be to fuse a riboswitch receptor domain to a self-cleaving ribozyme such that ligand binding to the receptor stabilizes the ribozyme structure and activates it for cleavage. This approach has been successfully used to make designer molecular switches in which an engineered RNA receptor is coupled to a self-cleaving ribozyme[49]. Thus, almost any class of natural riboswitch receptors should be amenable to the same technology. Although rational design strategies might be a more productive route to new riboswitch-binding compounds, the availability of high-throughput screening methods for riboswitches means that researchers seeking to develop distinct riboswitch-targeting classes of compounds are equipped with some of the same tools as those working with protein targets.

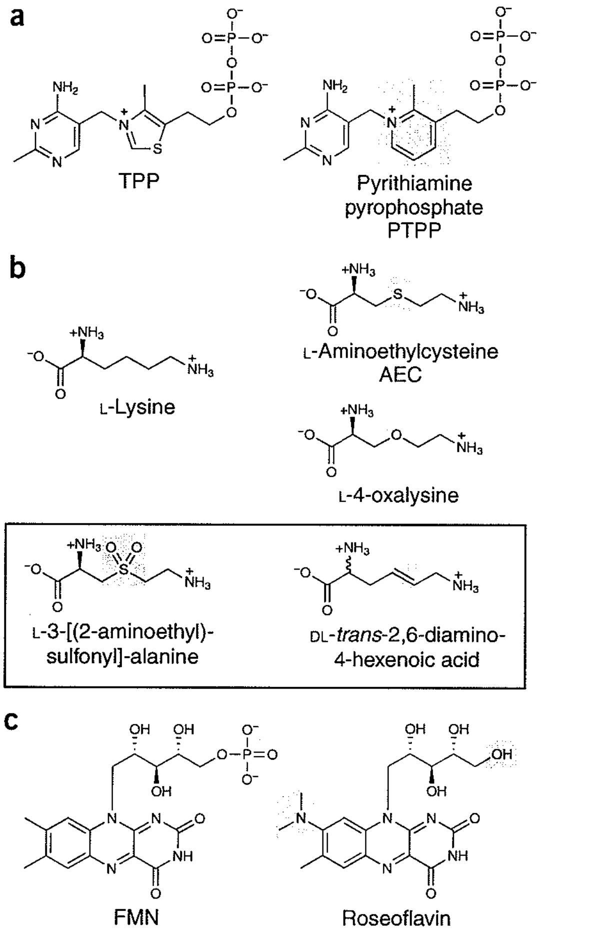


Figure 4 The chemical structures of riboswitch-binding natural metabolites (left) and antibacterial metabolite analogs (right). (**a**) Compounds that bind TPP riboswitches. (**b**) Compounds that bind lysine riboswitches. (**c**) Compounds that bind FMN riboswitches. Shaded regions highlight the functional groups that differ from the natural metabolite. The expected protonation state under physiological conditions is shown for each functional group. The inset in **b** shows recently identified antibacterial lysine analogs that bind to lysine riboswitches.

Future questions for riboswitch-targeting antibacterials

Although riboswitch-targeting compounds have been discovered that inhibit bacterial growth *in vitro*, some questions remain about whether these or similar compounds could cure an infection in a clinical setting. A principal question is the degree to which a riboswitch-targeting compound can repress gene expression. Although genetic experiments have identified riboswitch-regulated genes whose deletion prevents growth or virulence, it is unclear whether targeting the riboswitch for such genes will mimic a genetic knockout. In one case, a class 1 SAM-binding riboswitch in *Bacillus clausii* appears to repress the synthesis of the adjacent ORF to less than one transcript per cell at high ligand concentrations[50]. Similarly, at high lysine concentrations, a lysine riboswitch in *B. subtilis* completely represses reporter gene expression[34,35]. In contrast, a recent microarray analysis suggested that the expression of FMN riboswitch–regulated transcripts in *B. subtilis* is only modestly reduced when cells are grown in high concentrations of riboflavin[51] (riboflavin is phosphorylated to FMN inside bacteria). It could be that a fraction of the riboswitch receptors folds into an alternate inactive structure that cannot repress expression, even at high FMN concentrations. Even so, it is possible that partial repression might still be enough to halt growth. For antibacterial drug discovery it will be important to identify the riboswitches that can repress the expression of a crucial gene below a level that is required for survival.

Even in cases where metabolite biosynthesis can be completely repressed by riboswitch targeting, the bacteria might still grow if the metabolite can be imported from the host tissue. For example, although a riboswitch-targeting compound can presumably prevent lysine biosynthesis in *B. anthracis* by completely repressing *lysA* and *lysC*, the bacteria could still import lysine from their environment through a protein whose expression is not regulated by a riboswitch[52]. Accordingly, lysine riboswitch–targeting compounds that prevent *Bacillus* species from growing in media without lysine do not inhibit *B. anthracis* in rich media[34]. A more attractive class of targets might be the FMN riboswitches found in Gram-negative pathogens. Because these organisms cannot transport riboflavin, they are dependent upon riboflavin biosynthesis[53], which in many cases is regulated by an FMN riboswitch. Thus, a drug that can repress riboflavin biosynthesis by targeting this riboswitch should be lethal. Ultimately, it will be important to choose riboswitch targets that regulate the biosynthesis of metabolites that are not readily available from mammalian tissues or that cannot be transported by the pathogen.

Riboswitches examined to date have K_D values ranging from the mid-picomolar[50] to mid-micromolar[46] range. However, for some riboswitches, the speed of ligand association rather than the ligand-binding affinity controls whether the adjoining gene is expressed[54,55]. The 5′-UTR of the *B. subtilis ribDEAHT* operon, which codes for five riboflavin biosynthesis genes, contains an FMN-binding riboswitch[41]. During transcription, the sequence immediately downstream of the FMN receptor folds into either a terminator or antiterminator hairpin, depending on the occupancy state of the receptor. To form a terminator, the ligand must bind to the receptor and stabilize P1 before the 3′-half of the anti-terminator hairpin is synthesized (analogous to GENE ON, **Fig. 1a**). Because RNA polymerase is very fast, this genetic 'decision' occurs long before the ligand-receptor interaction can reach thermodynamic equilibrium[54]. As a result, the genetic decision depends on the speed at which the ligand binds, rather than its affinity for the receptor. For this reason, screening assays would be best designed to report either the binding speed of each test compound or, ideally, the ability of each compound to cause repression.

The potential for toxicity of riboswitch-targeted antibacterials in humans is a major concern in drug design. It is possible that compounds that target riboswitches in pathogenic bacteria might cause similar effects in their human hosts if the host also carries that riboswitch. Although the TPP riboswitch class has been found in plants[8], neither this class nor any other riboswitch class has been discovered in humans. If this trend holds, then riboswitch-targeting compounds are unlikely to cause toxic effects by unintentionally binding to homologous gene-regulating RNAs in humans.

Because riboswitches recognize fundamental metabolites, mammalian cells most likely have proteins that could be competitively inhibited by a compound that resembles the natural metabolite. For example, in addition to binding to TPP riboswitches, pyrithiamine inhibits thiamine pyrophosphokinase from rats[56]. Similarly, AEC can be transported into mammalian cells, where it is incorporated into proteins[57]. Thus, like other antibiotic classes, riboswitch-targeting compounds will need to be designed, that do not inhibit mammalian proteins. This should be achievable, given that structural analyses suggest that certain metabolites are recognized differently by riboswitches and proteins. Specifically, TPP-binding riboswitches use two metal ions to coordinate the pyrophosphate moiety of TPP[25], whereas TPP-binding proteins use one metal ion[25]. Likewise, the conformation of SAM is very different when bound to a riboswitch than when bound to most SAM-binding proteins[27]. Based on these differences, it should be possible to rationally design riboswitch-targeting compounds that are not recognized by mammalian enzymes.

The evolution of bacterial resistance must be taken into consideration when developing any antibacterial drug. Thus far, very little data have addressed the evolution of resistance to riboswitch-targeting compounds. One common mechanism by which bacteria become resistant to known antibiotics is by expressing a protein that modifies or exports the drug. Some riboswitch-targeting compounds might be susceptible to this mechanism, since many bacteria already express proteins that act on natural metabolites. Mutations that cause overexpression of one of these proteins could confer resistance to a compound that resembles the natural metabolite. Indeed, *B. subtilis* can evolve resistance to the antibacterial action of 2-fluoroadenine by overexpressing a purine exporter[58]. Similarly, *B. subtilis* strains selected for pyrithiamine resistance often bear a mutation in the TPP riboswitch that regulates the *tenA* operon[9]. In addition to derepressing thiamine biosynthesis, this mutation causes the overexpression of the TenA protein—a putative thiaminase that could hydrolyze pyrithiamine[59]. Thus, as in the case of 2-fluoroadenine, bacteria can evolve resistance to pyrithiamine by overexpressing proteins already present in their genomes. With this in mind, riboswitch-targeting compounds should be designed that are as chemically dissimilar to the natural metabolite as possible, so as to minimize their susceptibility to this resistance mechanism.

Bacteria might also evolve resistance to riboswitch-targeting drugs through a mutation that disrupts binding to the riboswitch receptor. As discussed earlier, resistance to AEC, pyrithiamine and roseoflavin can emerge in a laboratory setting through mutation of a lysine, TPP or FMN riboswitch, respectively, although the frequencies of mutation have not yet been determined. It is unclear how general this mechanism would be for evolving resistance to riboswitch-targeting compounds. In cases where several riboswitches of the same class are targeted by a single compound, one might expect that a single point mutation would be insufficient to confer resistance. Rather, a mutation would need to occur in each riboswitch that regulates an important gene. However, these mutations would also disrupt the binding of the natural ligand, resulting in deregulation of the associated biosynthesis pathway(s). This could impair survival of the bacteria, especially in an infectious setting. Indeed, a mutation in *B. subtilis* that deregulates the expression of the riboswitch-regulated SAM synthetase gene impairs growth in laboratory cultures[60]. Perhaps overexpression of SAM synthetase depletes bacteria of methionine or causes accumulation of *S*-adenosylhomocysteine (SAH), a toxic demethylated product of SAM. Regardless, this example illustrates that some riboswitch mutations that deregulate biosynthesis could impair bacterial growth. Thus, in some cases, it may be difficult for a pathogen to evolve resistance to a riboswitch-targeting compound by a point mutation in the receptor. Clearly, many questions must still be answered about the susceptibility of riboswitch-targeting antibacterials to this and other resistance mechanisms.

Conclusions

As with any new drug target, researchers seeking to develop riboswitch-targeting antibiotics may face some challenges. Even so, the body of work thus far underscores the promise for riboswitch drug development. The discovery that known antibacterial compounds may function by targeting a riboswitch establishes the feasibility of the approach. The complexity and selectivity of the ligand-binding domain of riboswitches should allow highly active compounds to be designed that target riboswitches but not other cellular RNAs or proteins. To identify such compounds, researchers can use much of the same technology that has been used to discover antibacterial drugs directed at other targets. Ultimately, the existence of multiple riboswitch classes in a variety of bacterial pathogens offers hope that several distinct classes of new drugs could be developed to help replenish the arsenal of antibiotic compounds.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests (see the *Nature Biotechnology* website for details).

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Theuretzbacher, U. & Toney, J.H. Nature's clarion call of antibacterial resistance: are we listening? *Curr. Opin. Investig. Drugs* **7**, 158–166 (2006).
2. D'Costa, V.M., McGrann, K.M., Hughes, D.W. & Wright, G.D. Sampling the antibiotic resistome. *Science* **311**, 374–377 (2006).
3. Mandal, M. & Breaker, R.R. Gene regulation by riboswitches. *Nat. Rev. Mol. Cell Biol.* **5**, 451–463 (2004).
4. Tucker, B.J. & Breaker, R.R. Riboswitches as versatile gene control elements. *Curr. Opin. Struct. Biol.* **15**, 342–348 (2005).
5. Winkler, W.C. & Breaker, R.R. Regulation of bacterial gene expression by riboswitches. *Annu. Rev. Microbiol.* **59**, 487–517 (2005).
6. Winkler, W., Nahvi, A. & Breaker, R.R. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature* **419**, 952–956 (2002).
7. Rodionov, D.A., Vitreschak, A.G., Mironov, A.A. & Gelfand, M.S. Comparative genomics of thiamin biosynthesis in procaryotes. New genes and regulatory mechanisms. *J. Biol. Chem.* **277**, 48949–48959 (2002).
8. Sudarsan, N., Barrick, J.E. & Breaker, R.R. Metabolite-binding RNA domains are present in the genes of eukaryotes. *RNA* **9**, 644–647 (2003).
9. Sudarsan, N., Cohen-Chalamish, S., Nakamura, S., Emilsson G.M. & Breaker, R.R. Thiamine pyrophosphate riboswitches are targets for the antimicrobial compound pyrithiamine. *Chem. Biol.* **12**, 1325–1335 (2005).
10. Zhang, R., Ou, H.-Y. & Zhang, C.-T. DEG, a database of essential genes. *Nucleic Acids Res.* **32**, D271–D272 (2004).
11. Forsyth, R.A. *et al.* A genome-wide strategy for the identification of essential genes in *Staphylococcus aureus*. *Mol. Microbiol.* **43**, 1387–1400 (2002).
12. Dunman, P.M. *et al.* Transcription profiling-based identification of *Staphylococcus aureus* genes regulated by the *agr* and/or *sarA* loci. *J. Bacteriol.* **183**, 7341–7353 (2001).
13. Epshtein, V., Mironov, A.S. & Nudler, E. The riboswitch-mediated control of sulfur metabolism in bacteria. *Proc. Natl. Acad. Sci. USA* **100**, 5052–5056 (2003).
14. Murphy, B.A., Grundy, F.J. & Henkin, T.M. Prediction of gene function in methylthioadenosine recycling from regulatory signals. *J. Bacteriol.* **184**, 2314–2318 (2002).
15. Winkler, W.C., Nahvi, A., Sudarsan, N., Barrick, J.E. & Breaker, R.R. An mRNA structure that controls gene expression by binding *S*-adenosylmethionine. *Nat. Struct. Biol.* **10**, 701–707 (2003).
16. Nahvi, A., Barrick, J.E. & Breaker, R.R. Coenzyme B$_{12}$ riboswitches are widespread genetic control elements in prokaryotes. *Nucleic Acids Res.* **32**, 143–150 (2004).
17. Monaghan, R.L. & Barrett, J.F. Antibacterial drug discovery-then, now and the genomics future. *Biochem. Pharmacol.* **71**, 901–909 (2006).
18. Knowles, D.J., Foloppe, N., Matassova, N.B. & Murchie, A.I. The bacterial ribosome, a promising focus for structure-based drug design. *Curr. Opin. Pharmacol.* **2**, 501–506 (2002).
19. Steitz, T.A. On the structural basis of peptide-bond formation and antibiotic resistance from atomic structures of the large ribosomal subunit. *FEBS Lett.* **579**, 955–958 (2005).
20. Zaman, G.J.R. & Michiels, P.J.A. Targeting RNA with small molecule drugs in *Trends in RNA Research* (ed. McNamara, P.) 1–21 (Nova Science Publishers, Inc., Hauppauge, NY, 2006).
21. Hermann, T. & Tor, Y. RNA as a target for small-molecule therapeutics. *Expert. Opin. Ther. Pat.* **15**, 49–62 (2005).
22. Mandal, M., Boese, B., Barrick, J.E., Winkler, W.C. & Breaker, R.R. Riboswitches control fundamental biochemical pathways in *Bacillus subtilis* and other bacteria. *Cell* **113**, 577–586 (2003).
23. Serganov, A. *et al.* Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs. *Chem. Biol.* **11**, 1729–1741 (2004).
24. Batey, R.T., Gilbert, S.D. & Montange, R.K. Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine. *Nature* **432**, 411–415 (2004).
25. Serganov, A., Polonskaia, A., Phan, A.T., Breaker, R.R. & Patel, D.J. Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch. *Nature* **441**, 1167–1171 (2006).
26. Thore, S., Leibundgut, M. & Ban, N. Structure of the eukaryotic thiamine pyrophosphate riboswitch with its regulatory ligand. *Science* **312**, 1208–1211 (2006).
27. Montange, R.K. & Batey, R.T. Structure of the *S*-adenosylmethionine riboswitch regulatory mRNA element. *Nature* **441**, 1172–1175 (2006).
28. Lim, J., Winkler, W.C., Nakamura, S., Scott, V. & Breaker, R.R. Molecular-recognition characteristics of SAM-binding riboswitches. *Angew. Chem. Int. Edn Engl.* **45**, 964–968 (2006).
29. Robbins, W.J. The pyridine analog of thiamine and the growth of fungi. *Proc. Natl. Acad. Sci. USA* **27**, 419–422 (1941).
30. Woolley, D.W. & White, A.G.C. Selective reversible inhibition of microbial growth with pyrithiamine. *J. Exp. Med.* **78**, 489–497 (1943).
31. Iwashima, A., Wakabayashi, Y. & Nose, Y. Formation of thiamine pyrophosphate in brain tissue. *J. Biochem.* **79**, 845–847 (1976).
32. Shiota, T., Folk, J.E. & Tietze, F. Inhibition of lysine utilization in bacteria by *S*-(beta-aminoethyl) cysteine and its reversal by lysine peptides. *Arch. Biochem. Biophys.* **77**, 372–377 (1958).
33. McCord, T., Ravel, J., Skinner, C. & Shive, W. DL-4-oxalysine, an inhibitory analog of lysine. *J. Am. Chem. Soc.* **79**, 5693–5696 (1957).
34. Blount, K.F., Wang, X.J., Lim, J., Sudarsan, N. & Breaker, R.R. Antibacterial compounds that target lysine riboswitches. *Nat. Chem. Biol.* in the press (2007).
35. Sudarsan, N., Wickiser, J.K., Nakamura, S., Ebert, M.S. & Breaker, R.R. An mRNA structure in bacteria that controls gene expression by binding lysine. *Genes Dev.* **17**, 2688–2697 (2003).
36. Vold, B., Szulmajster, J. & Carbone, A. Regulation of dihydropicolinate synthase and aspartate kinase in *Bacillus subtilis*. *J. Bacteriol.* **121**, 970–974 (1975).
37. Lu, Y., Shevtchenko, T. & Paulus, H. Fine structure mapping of *cis*-acting control sites in the *lysC* operon of *Bacillus subtilis*. *FEMS Microbiol. Lett.* **71**, 23–27 (1992).
38. Matsui, K. *et al.* Riboflavin production by roseoflavin-resistant strains of some bacteria. *Agric. Biol. Chem.* **46**, 2003–2008 (1982).
39. Berezovskii, V.M., Stepanov, A.I., Polyakova, N.A., Tulchinskaya, L.S. & Kukanova, A.Y. Studies of a group of allo- and isoallxazine. XLVI. Synthesis and biological specificity of amino analogs. *Bioorg. Khim,* **3**, 521–524 (1977).
40. Gelfand, M.S., Mironov, A.A., Jomantas, J., Kozlov, Y.I. & Perumov, D.A. A conserved RNA structure element involved in the regulation of bacterial riboflavin synthesis genes. *Trends Genet.* **15**, 439–442 (1999).
41. Winkler, W.C., Cohen-Chalamish, S. & Breaker, R.R. An mRNA structure that controls gene expression by binding FMN. *Proc. Natl. Acad. Sci. USA* **99**, 15908–15913 (2002).
42. Burgess, C., O'Connel-Motherway, M., Sybesma, W., Hugenholtz, J. & van Sinderen, D. Riboflavin production in *Lactococcus lactis*: potential for *in situ* production of vitamin-enriched foods. *Appl. Environ. Microbiol.* **70**, 5769–5777 (2004).
43. Kreneva, R.A. & Perumov, D.A. Genetic mapping of regulatory mutations of *Bacillus subtilis* riboflavin operon. *Mol. Gen. Genet.* **222**, 467–469 (1990).
44. Kil, Y.V., Mironov, V.N., Gorishin, I., Kreneva, R.A. & Perumov, D.A. Riboflavin operon of *Bacillus subtilis*: unusual symmetric arrangement of the regulatory region. *Mol. Gen. Genet.* **233**, 483–486 (1992).
45. Edwards, T.E. & Ferré-D'Amaré, A.R. Crystal structures of the thi-box riboswitch bound to thiamine pyrophosphate analogs reveal adaptive RNA-small molecule recognition. *Structure* **14**, 1459–1468 (2006).
46. Winkler, W.C., Nahvi, A., Roth, A., Collins, J.A. & Breaker, R.R. Control of gene expression by a natural metabolite-responsive ribozyme. *Nature* **428**, 281–286 (2004).
47. Mayer, G. & Famulok, M. High-throughput-compatible assay for *glmS* riboswitch metabolite dependence. *ChemBioChem* **7**, 602–604 (2006).
48. Blount, K.F., Puskarz, I., Penchovsky, R. & Breaker, R.R. Development and application of a high-throughput assay for *glmS* riboswitch activators. *RNA Biology* **3**, 77–81 (2006).
49. Seetharaman, S., Zivarts, M., Sudarsan, N. & Breaker, R.R. Immobilized RNA switches for the analysis of complex chemical and biological mixtures. *Nat. Biotechnol.* **19**, 336–341 (2001).
50. Sudarsan, N., Hammond, M.C., Block, K., Welz, R. & Breaker, R.R. Tandem riboswitch architectures exhibit complex gene control functions. *Science* **314**, 300–304 (2006).
51. Lee, J.-M. *et al.* RNA expression analysis using an antisense *Bacillus subtilis* genome array. *J. Bacteriol.* **183**, 7371–7380 (2001).
52. Rodionov, D.A., Vitreschak, A.G., Mironov, A.A. & Gelfand, M.S. Regulation of lysine biosynthesis and transport genes in bacteria: yet another RNA riboswitch? *Nucleic Acids Res.* **31**, 6748–6757 (2003).
53. Bacher, A., Eberhardt, S., Fischer, M., Kis, K. & Richter, G. Biosynthesis of vitamin B$_2$ (riboflavin). *Annu. Rev. Nutr.* **20**, 153–167 (2000).
54. Wickiser, J.K., Winkler, W.C., Breaker, R.R. & Crothers, D.M. The speed of RNA transcription and metabolite binding kinetics operate an FMN riboswitch. *Mol. Cell* **18**, 49–60 (2005).
55. Wickiser, J.K., Cheah, M.T., Breaker, R.R. & Crothers, D.M. The kinetics of ligand binding by an adenine-sensing riboswitch. *Biochemistry* **44**, 13404–13414 (2005).
56. Koedam, J.C. The mode of action of pyrithiamine as an inducer of thiamine deficiency. *Biochim. Biophys. Acta* **29**, 333–344 (1958).
57. Di Girolamo, M., Di Girolamo, A., Cini, C., Coccia, R. & De Marco, C. Thialysine utilization for protein synthesis by CHO cells. *Physiol. Chem. Phys. Med. NMR* **18**, 159–164 (1986).
58. Saxild, H.H. & Nygaard, P. Genetic and physiological characterization of *Bacillus subtilis* mutants resistant to purine analogs. *J. Bacteriol.* **169**, 2977–2983 (1987).
59. Haas, A.L., Laun, N.P. & Begley, T.P. Thi20, a remarkable enzyme from *Saccharomyces cerevisiae* with dual thiamine biosynthetic and degradation activities. *Bioorg. Chem.* **33**, 338–344 (2005).
60. McDaniel, B.A., Grundy, F.J., Kurlekar, V.P., Tomsic, J. & Henkin, T.M. Identification of a mutation in the *Bacillus subtilis S*-adenosylmethionine synthetase gene that results in derepression of S-box gene expression. *J. Bacteriol.* **188**, 3674–3681 (2006).
61. Barrick, J.E. *et al.* New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proc. Natl. Acad. Sci. USA* **101**, 6421–6426 (2004).
62. Corbino, K.A. *et al.* Evidence for a second class of *S*-adenosylmethionine riboswitches and other regulatory RNA motifs in alpha-proteobacteria. *Genome Biol.* **6**, R70.1–R70.10 (2005).
63. Mandal, M. *et al.* A glycine-dependent riboswitch that uses cooperative binding to control gene expression. *Science* **306**, 275–279 (2004).

Guanine riboswitch variants from *Mesoplasma florum* selectively recognize 2′-deoxyguanosine

Jane N. Kim[†], Adam Roth[‡], and Ronald R. Breaker[†‡§¶]

[†]Department of Molecular, Cellular, and Developmental Biology, [§]Department of Molecular Biophysics and Biochemistry, and [‡]Howard Hughes Medical Institute, Yale University, P.O. Box 208103, New Haven, CT 06520-8103

Edited by Jeffrey W. Roberts, Cornell University, Ithaca, NY, and approved August 21, 2007 (received for review June 22, 2007)

Several mRNA aptamers have been identified in *Mesoplasma florum* that have sequence and structural features resembling those of guanine and adenine riboswitches. Two features distinguish these RNAs from established purine-sensing riboswitches. All possess shortened hairpin-loop sequences expected to alter tertiary contacts known to be critical for aptamer folding. The RNAs also carry nucleotide changes in the core of each aptamer that otherwise is strictly conserved in guanine and adenine riboswitches. Some aptamers retain the ability to selectively bind guanine or adenine despite these mutations. However, one variant type exhibits selective and high-affinity binding of 2′-deoxyguanosine, which is consistent with its occurrence in the 5′ untranslated region of an operon containing ribonucleotide reductase genes. The identification of riboswitch variants that bind nucleosides and reject nucleobases reveals that natural metabolite-sensing RNA motifs can accrue mutations that expand the diversity of ligand detection in bacteria.

allosteric RNA | aptamer | metabolite | ribonucleotide reductase | transcription termination

Gene-control elements called riboswitches (1) are mRNA motifs typically found in the 5′ untranslated regions of bacterial mRNAs. Riboswitches selectively bind small molecules, and structural changes within the 5′ untranslated regions are usually harnessed to control the expression of the adjoining ORF. The architectures of these RNAs are commonly formed from a metabolite-binding aptamer domain and an expression platform (2–5), although more diverse assemblies of aptamers and expression platforms have been found that yield more complex gene-control characteristics (6–9).

Each aptamer adopts a complex secondary- and tertiary-structured fold to form a conserved receptor for the ligand (10–17). This demand for precise structure formation and specific molecular recognition causes the aptamer domains to be highly conserved even among distantly related species. In contrast, the expression platform can adopt a variety of different structures provided it maintains its responsiveness to the occupation state of the aptamer domain.

Computer-aided searches based on conserved RNA sequences and structures have been used to identify representatives of numerous riboswitch classes (18–24). However, these bioinformatics algorithms can fail to identify variants of known riboswitch classes that differ substantially from the established aptamer consensus. Furthermore, there could be exceedingly rare classes of riboswitch aptamers or exceptionally small aptamers that will be missed by existing bioinformatics algorithms because there are too few representatives for comparison or they have too few conserved features. Given these limitations, many new riboswitch classes might remain undiscovered, and the true number of metabolite-sensing aptamer folds could greatly exceed the many types published to date (25).

Indications that there are many variant, small, or rare riboswitch classes to be discovered come from several recent reports of new riboswitch classes. For example, a single C-to-U mutation within the core of guanine riboswitches can change the specificity of ligand binding to adenine (11, 26–30). Also, two related types of riboswitch aptamers for the modified nucleobase 7-aminomethyl-7-deazaguanine have been identified that require as few as 34 nucleotides to form a selective and high-affinity binding pocket (31). Moreover, there are three classes of aptamers for *S*-adenosylmethionine (SAM) (23, 24, 32) whose representatives are more rare in bacteria than the SAM-I class of riboswitches commonly found in Gram-positive bacteria (33–35). These findings indicate that a far greater diversity of metabolite-sensing riboswitches exists that might be difficult for existing search strategies to definitively identify and classify.

One possibility is that some organisms will have recently evolved riboswitch classes with aptamers that are unique in architecture or ligand specificity. If such boutique riboswitches can easily emerge through evolution, there could be far more riboswitches than have been discovered to date. In this report, we describe a series of aptamers that are exceedingly rare among sequenced genomes and have been identified only in the bacterial species *Mesoplasma florum*. One subclass of these RNAs is selective for 2′-deoxyguanosine (2′-dG). Our findings highlight the capacity for metabolite-binding RNAs to evolve specificities toward structurally related derivatives and further demonstrate that exceptionally rare riboswitch classes are likely to be present in some organisms.

Results and Discussion

Consensus Sequences and Structures of Guanine- and Adenine-Sensing Riboswitches. Guanine-sensing riboswitches usually reside upstream of genes involved in purine biosynthesis, salvage, and transport (26). The guanine riboswitch aptamer from the *Bacillus subtilis xpt-pbuX* mRNA exhibits a K_d for guanine of ≈5 nM. Ligand binding to this aptamer causes transcription termination, and a similar gene-control mechanism is predicted for most other guanine riboswitches as well. X-ray crystallography has been performed on this aptamer bound to either guanine or the functional analog hypoxanthine (10, 11). In both instances, the ligands are almost completely enveloped by the RNA. Similarly, a related aptamer that binds adenine by using a similar architecture also engulfs the ligand (11).

The tight ligand-binding pocket of this aptamer class is formed by conserved nucleotides at the junction of three stems termed P1, P2, and P3 (Fig. 1*A*). When the ligand is bound, the aptamer adopts a conformation with the P2 and P3 stems extending parallel to one another. This structure is held in place by Watson–Crick base-pairing interactions and other hydrogen bonds formed between the loops of these stems, called L2 and L3. Most of the highly conserved nucleotides forming the ligand-binding core of the aptamer are

Author contributions: J.N.K., A.R., and R.R.B. designed research; J.N.K. and A.R. performed research; J.N.K., A.R., and R.R.B. analyzed data; and J.N.K., A.R., and R.R.B. wrote the paper.

Conflict of interest statement: R.R.B. is a cofounder of BioRelix, a biotechnology company that has licensed riboswitch technology from Yale University for antibiotics development.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Abbreviations: 2′-dA, 2′-deoxyadenosine; 2′-dG, 2′-deoxyguanosine; 2′-d-2,6-DAP, 2′-deoxy-2,6-diathinopurine nucleoside.

[¶]To whom correspondence should be addressed. E-mail: ronald.breaker@yale.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0705884104

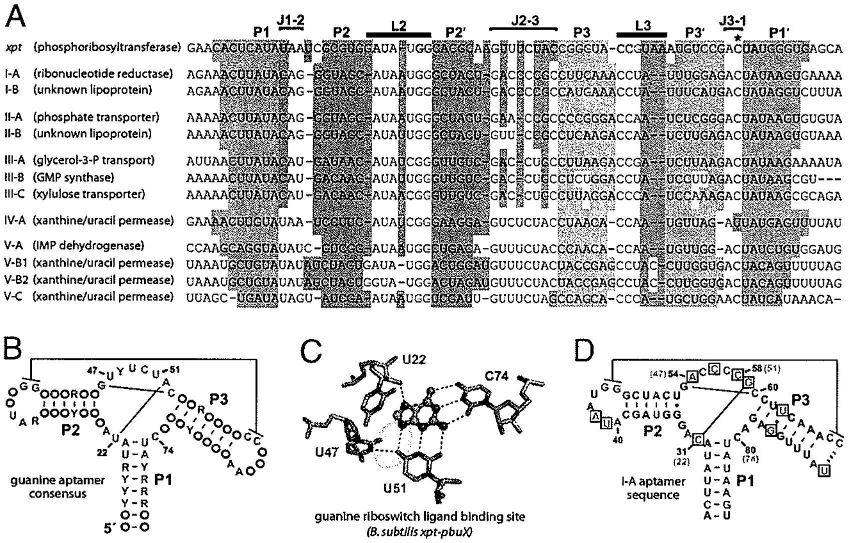

A
P1
J1-2
P2
L2
P2′
J2-3
P3
L3
P3′
J3-1
*
P1′
xpt (phosphoribosyltransferase) GAACACUCAUAUAAUCGCGUGGAUA-UGGCACGCAAGUUUCUACCGGGUA-CCGUAAAUGUCCGACUAUGGGUGAGCA
I-A (ribonucleotide reductase) AGAAACUUAUACAG-GGUAGC-AUAAUGGGCUACU-GACGCCGCCUUCAAACCUA--UUUGGAGACUAUAAGUGAAAA
I-B (unknown lipoprotein) AGAAACUUAUACAG-GGUAGC-AUAAUGGGCUACU-GACGCCGCCAUGAAACCUA--UUUCAUGACUAUAGGUCUUUA
II-A (phosphate transporter) AAAAACUUAUACAG-GGUAGC-AUAAUGGGCUACU-GAA-CCGCCCCGGGACCAA--UCUCGGGACUAUAAGUGUGUA
II-B (unknown lipoprotein) AAAAACUUAUACAG-GGUAGC-AUAUUGGGCUACU-GUU-CCGCCUCAAGACCAA--UCUUGAGACUAUAAGUGUAAA
III-A (glycerol-3-P transport) AUUAACUUAUACAU-GAUAAC-AUAUCGGGUUGUC-GAC-CUGCCUUAAGACCGA--UCUUAAGACUAUAAGAAAAUA
III-B (GMP synthase) AAAAACUUAUACAU-GACAAC-AUAUUGGGUUGUC-GAC-CUGCCUCUGGACCUA--UCCUUAGACUAUAAGCGU---
III-C (xylulose transporter) AAAAACUUAUACAU-GACAAC-AUAACGGGUUGUC-GAC-CUGCCUUAGGACCCA--UCCAAAGACUAUAAGCGCAGA
IV-A (xanthine/uracil permease) GAAAACUUGUAUAA-UCCUUC-AUAUCGGGAAGGA-GUCUCUACCUAACA-CCAA--UGUUAG-AUUAUGACUUUUAU
V-A (IMP dehydrogenase) CCAAGCAGGUAUAUC-GUCGG-AUAAUGGCUGACA-GUUUCUACCCAACA-CCAA--UGUUGG-ACUAUCUGUGGAUG
V-B1 (xanthine/uracil permease) UAAAUGCUGUAUAUAUCUAGUGAUA-UGGACUGGAUGUUUCUACUACCGAGCCUAC-CUUGGUGACUACAGUUUUUAG
V-B2 (xanthine/uracil permease) UAAAUGCUGUAUAUAUCUAGUGGUA-UGGACUAGAUGUUUCUACUACCGAGCCUAC-CUUGGUGACUACAGUUUUUAG
V-C (xanthine/uracil permease) UUAGC-UGAUAUAGU-AUCGA-AUAAUGGUCGAUU-GUUUCUAGCCAGCA-CCCA--UGCUGGAACUAUCAUAAACA-
B
P2
P3
P1
22
74
47
51
guanine aptamer consensus
5′
C
U22
C74
U47
U51
guanine riboswitch ligand binding site
(B. subtilis xpt-pbuX)
D
(47) 54
58 (51)
60
P2
P3
P1
40
31 (22)
80 (74)
I-A aptamer sequence

Fig. 1. Sequence and structural features of guanine riboswitch aptamers and several newly found RNAs. (*A*) Sequence alignment comparing the *xpt* guanine riboswitch aptamer sequence from the *xpt-pbuX* mRNA from *B. subtilis* with related sequences from *Mesoplasma florum* (types I, II, III, and IV) and from *Oenococcus oeni, Vibrio sp., Vibrio splendidus,* and *Leuconostoc mesenteroides* (types V-A, V-B1, V-B2, and V-C, respectively). The known or putative functions of the genes immediately downstream of each sequence are noted as predicted elsewhere (54). Nucleotides corresponding to pairing regions P1, P2, and P3 are shaded blue, green, and orange, respectively. Nucleotides corresponding to loop regions (L) and joining regions (J) also are identified. The asterisk identifies the C nucleotide in *xpt* that forms a Watson–Crick base pair with the guanine ligand. Nucleotides shaded gray are mutated relative to the highly conserved nucleotides denoted in red that are typically found in the J and L regions of guanine and adenine riboswitches or nucleotides that are inserted or deleted in these regions. (*B*) Consensus sequence and secondary structure of guanine riboswitch aptamers. Nucleotides in red are present in >90% of the known representatives. Circles identify nucleotides whose base identities are not conserved, and lines indicate Watson–Crick base pairing. Nucleotides that form hydrogen-bonding interactions with the guanine ligand are identified according to the numbering system used previously for the *xpt* aptamer (10, 11, 26). (*C*) Structural model of the guanine-binding site formed by the *xpt* aptamer docked to guanine (10, 11). Dashed lines identify hydrogen-bonding contacts between the aptamer nucleotides (numbered as described in *B*) and the ligand. The shaded area identifies the space that would be occupied by the sugar moiety of a guanine nucleoside. (*D*) Sequence and secondary structure of the type I-A aptamer from *M. florum*. Boxed nucleotides depicted in blue identify variations from the *xpt* aptamer that occur at otherwise highly conserved positions. Dashed line represents a 3-nucleotide deletion compared with the L3 sequence of *xpt*. Nucleotide numbers are as described in Fig. 2, with the equivalent positions for the *xpt* aptamer depicted in parentheses. Other notations are as defined for *B*.

present in joining regions J1-2, J2-3, and J3-1, which link the three stems together (Fig. 1*B*).

The joining regions carry four nucleotides that form hydrogen bonds with functional groups of the purine ligand (Fig. 1*C*). One key interaction is made by nucleotide C74 of the *xpt* aptamer, which forms a Watson–Crick base pair with guanine. Interestingly, several variants of this RNA motif were found that carry a C-to-U mutation at the equivalent position in the structure, and these RNAs reject guanine and bind adenine with affinities measured in the mid nanomolar range (27, 28). An atomic-resolution model of an adenine riboswitch aptamer from the *add* gene in *B. subtilis* confirmed that adenine forms a Watson–Crick base-pairing interaction with the variant U nucleotide, and that other features typical of guanine riboswitches remained essentially identical (11). Indeed, this single-nucleotide change at position 74 is sufficient to change the specificity of guanine aptamers to adenine and vice versa.

Discovering More Distant Homologs of Guanine and Adenine Riboswitches. We used a bioinformatics search strategy to discover variant purine riboswitch candidates. This process was achieved with an algorithm that identifies sequences that closely correspond to the consensus sequence and secondary structure features of known purine-sensing aptamers (see *Materials and Methods*). The parameters of this search were set to allow recovery of low-quality matches, and we focused the most attention on sequences that deviate substantially from the consensus, but that nonetheless exist in genomic contexts consistent with riboswitch function.

We noticed one sequence in *M. florum* that could be threaded to conform reasonably well to the consensus structure, but that deviated in sequence at several key positions in the core and in loops L2 and L3. Despite these significant differences, the location of this sequence in the apparent 5′ untranslated region of the *guaAB* operon suggested that it might function as a purine-sensing riboswitch. A BLAST search for related sequences uncovered several more examples of this motif in *M. florum*. This bacterium is a nonparasitic member of the class Mollicutes, and organisms of this class are notable for their simplified cell structures and small genomes (36).

One of the *M. florum* sequences differed from characterized purine riboswitches only in the L2 and L3 regions, with the joining regions otherwise adhering to the consensus. To determine whether there might be other variant purine aptamers analogous to this RNA, we manually inspected sequences generated from the original search, scanning for sequence or structure irregularities. Three additional riboswitches were identified that contained shortened L3 sequences relative to the consensus, in addition to the eight *M. florum* examples mentioned previously.

Subsequently, we performed automated searches by using algorithms trained on all known purine riboswitches, as well as algorithms trained more narrowly on the variant sequences, but no additional variants were identified. In total, 12 new putative riboswitch examples were found in bacteria (Fig. 1*A*), all of which bear close similarity to the consensus sequence and structure established for guanine aptamers. These RNAs were classified into five types (I–V) based on the mutations they carry relative to the guanine aptamer consensus. Interestingly, eight of the RNAs representing types I-IV are present in *M. florum*.

One of the eight RNAs (IV-A) carries a C-to-U mutation at the position equivalent to nucleotide 74 of the *xpt* riboswitch in *B. subtilis*. Therefore, this RNA was predicted to sense adenine. The seven remaining RNAs from *M. florum* carry mutations at two or more positions that are highly conserved among known guanine and adenine riboswitch representatives (Fig. 1*A*). All seven RNAs carry mutations in otherwise conserved nucleotides in J1-2, J2-3, and L2 of the aptamer, and they also carry an L3 loop that is four nucleotides, rather than the seven or eight nucleotides normally present in known guanine-binding aptamers. The four remaining RNAs, classified as type V, are found in other bacterial species and carry the distinctive nucleotide changes in L3 and, in some instances, L2. Specific interactions between the L2 and L3 regions are known to be important for folding and function of guanine and adenine riboswitches (29), and therefore the loop mutations likely cause the variant RNAs to adopt a different structure for this tertiary interaction.

We speculated that the aptamer core mutations in the *M. florum* RNAs might substantially alter the ligand-binding pocket of each riboswitch, allowing it to recognize a metabolite other than guanine. For example, RNA I-A carries 39 nucleotide changes (including insertions and deletions) relative to the *xpt* RNA (Fig. 1*A*), and 10 of these changes occur at positions with nucleotide identities that are conserved in >90% of the known guanine riboswitch aptamers (Fig. 1*D*). In addition, three of the four nucleotides known to contact guanine in the *xpt* aptamer (nucleotides 22, 47, and 51) (Fig. 1 *B* and *C*) are mutated in the I-A (nucleotides 31, 54, and 58) (Fig. 1*D*) and I-B aptamers. Although the nucleotide corresponding to C74 of *xpt* in the I-A and I-B aptamers could retain its recognition of the base-pairing face of guanine, the other core mutations likely recognize other portions of a guanine-containing ligand. Moreover, these core mutations typically convert A and U residues to G and C residues, despite the fact that the *M. florum* genome has only 27% GC content. The acquisition of additional G and C residues in some variant aptamers suggests adaptation to a new function.

A Guanine Riboswitch Variant Binds 2′-dG. Frequently, the metabolite that is sensed by a riboswitch can be discerned by noting the function of the protein product of the downstream gene. Some of the RNA motifs are located upstream of either unannotated genes or genes with functions that appear to be unrelated to purine metabolism, and thus did not provide clues for possible ligands. However, some reside upstream of genes involved in purine biosynthesis or transport (Fig. 1*A*), suggesting that the variant RNAs bind guanine or a ligand that includes this nucleobase. Of particular interest was aptamer I-A, which resides upstream of genes encoding ribonucleotide reductase subunits. Ribonucleotide reductase enzymes convert ribonucleotides into their deoxyribonucleotide counterparts (37, 38). Given that three of the nucleotides mutated in the I-A aptamer are in the immediate vicinity of the N9 position of guanine in known guanine riboswitches (Fig. 1*C*), we speculated that this variant riboswitch might respond to 2′-dG or one of its 5′-phosphorylated derivatives.

In-line probing (39) was performed by using a series of guanine and guanosine derivatives (see *Materials and Methods* for a complete list) to determine the ligand specificity for all variant riboswitch types. In-line probing assays reveal shape changes in an aptamer that occur upon ligand binding. For example, I-A exhibits substantial structural modulation when 100 μM 2′-dG is present (Fig. 2*A*). Importantly, the pattern of spontaneous cleavage products is consistent with the formation of a three-stem junction similar to guanine and adenine aptamers (26–28), and the majority of the internucleotide linkages that become more structured upon 2′-dG addition (Fig. 2*B*) are in the predicted ligand-binding core of I-A RNA. Furthermore, in-line probing data collected at various concentrations of 2′-dG indicate changes in the extent of RNA cleavage at specific sites in the RNA that are consistent with a 1:1 binding of ligand with an apparent K_d of ≈80 nM (Fig. 2*C*).

Affinities and Specificities of Natural Variants of Guanine Riboswitch Aptamers. A previous study (26) revealed that the apparent K_d value of the *xpt* riboswitch aptamer for guanine is ≈5 nM. We conducted similar K_d determinations for the RNAs shown in Fig. 1*A* and expanded this process to include compounds similar to 2′-dG (Fig. 3 and data not shown). As expected, we find that the *xpt* RNA binds guanine most tightly, whereas the addition of a ribose or deoxyribose moiety on the N9 position of the purine ring causes a loss of binding affinity of nearly two orders of magnitude or more (Fig. 3*A Left*). Similarly, types III and V prefer binding guanine over various nucleoside derivatives by ≈100-fold or more. This observation is expected because these RNA types retain a U residue corresponding to nucleotide 51 in the *xpt* aptamer. The retention of this residue and a C at the position analogous to C74 (Fig. 1*C*) suggests that these RNAs can form at least six of the seven hydrogen bonds formed between known guanine aptamers and their guanine ligand. The type V RNAs retain the base identities for all ligand-binding nucleotides, and these RNAs exhibit K_d values that are most similar to those of the *xpt* aptamer for all of the ligands tested. Also as expected, the type IV RNA, which carries the C-to-U mutation at position 74, binds adenine more tightly than guanine and exhibits binding affinities for these ligands that are consistent with a previously studied adenine riboswitch aptamer (27, 28).

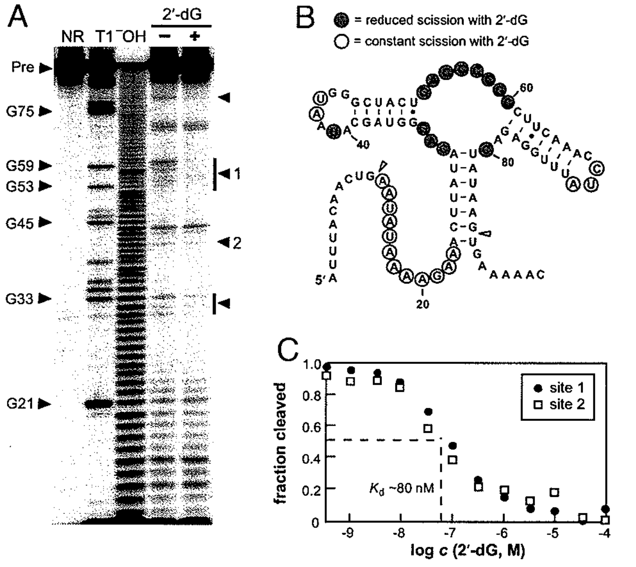


Fig. 2. Structural modulation of the I-A aptamer caused by binding of 2′-dG. (*A*) In-line probing analysis of the I-A aptamer. Fragmentation patterns of 5′ ^{32}P-labeled precursor RNAs (Pre) were established by incubating the RNAs in the absence (−) or presence (+) of 100 μM 2′-dG and separating spontaneous RNA cleavage products by using denaturing 6% PAGE. Lanes designated NR, T1, and $^-$OH identify RNA samples loaded after subjecting to no reaction, partial digestion with RNase T1, or partial digestion with alkali, respectively. Left arrowheads identify select bands corresponding to RNase T1 cleavage after G residues, and right arrowheads identify regions within the aptamer that undergo structural stabilization on ligand binding. Band intensity changes at sites 1 and 2 from a separate inline probing analysis (PAGE image not shown) were used to establish the apparent K_d as depicted in *C*. (*B*) Sites of structural flexibility and 2′-dG-mediated structural modulation. Data were derived from the image depicted in *A*. (*C*) Plot of the normalized fraction of RNA cleaved at sites 1 and 2 versus the logarithm of the concentration (c) of 2′-dG. The concentration of ligand required to cause half-maximal change in fraction cleaved (dashed lines) reflects the apparent K_d.

In contrast, type I RNAs most tightly bind 2′-dG relative to guanine and various nucleoside and nucleotide analogs (Fig. 3*A Right*). Most strikingly, both I-A and I-B RNAs discriminate by approximately two orders of magnitude against guanosine, which differs from 2′-dG by a single oxygen atom. This level of discrimination might be required by the cell to ensure that the expression of ribonucleotide reductase is controlled only by changing concentrations of a deoxyribonucleoside, and that expression is not inappropriately repressed by normal concentrations of the corresponding ribonucleoside. Similarly, type II RNAs exhibit the highest affinities for 2′-dG, but are far less selective for this compound than type I aptamers. Perhaps type II RNAs are intentionally less selective to permit changing concentrations of several guanine-containing compounds to modulate gene expression.

These findings also are consistent with the genomic arrangement of variant RNAs in *M. florum*. Specifically, aptamer types I-B, II-B, and III-A control individual genes located immediately adjacent to each other in the genome. If all three genes were controlled by only one compound, then a three-gene operon arrangement controlled by one riboswitch would be optimal. However, the presence of three types of aptamers next to these adjacent genes strongly suggests that their expression is under the control of riboswitches with three distinct specificities or affinities.

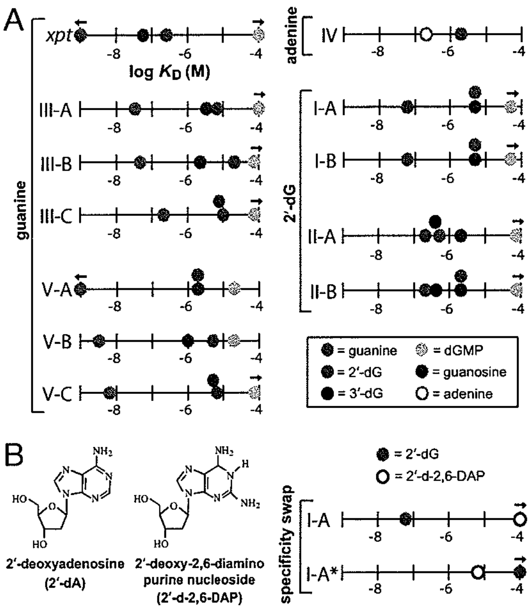


Fig. 3. Affinities and specificities of natural guanine aptamer variants. (*A*) Values for apparent K_d for various ligands were determined for the guanine riboswitch aptamer from the *B. subtilis xpt* RNA and with each variant aptamer type. An arrow above a circle indicates that the exact K_d value is either lower or higher than the scale of the plot. (*B*) Chemical structures of 2′-dA and the analog 2′-d-2,6-DAP and the apparent K_d values for 2′-dG and 2′-d-2,6-DAP with I-A RNA or a mutant version of this aptamer (I-A*) that carries a single C-to-U mutation at a position equivalent to nucleotide 74 (Fig. 1*B*). Compounds tested that failed to bind are not presented.

A Single Mutation Swaps Ligand Specificity of the 2′-dG Aptamer I-A. A C-to-U mutation at nucleotide 74 of the *xpt* aptamer can change its ligand specificity from guanine to adenine (27). This nucleotide forms a Watson–Crick base pair with the purine moiety, whereas other nucleotides make hydrogen-bonding interactions with other positions on the ligand (Fig. 1*C*) (10, 11, 40). Although most purine-sensing riboswitches respond to guanine, six natural examples of adenine-sensing riboswitches carrying a U at the equivalent position 74 have been identified (27, 41, 42), and a seventh example is represented by the type IV RNA reported in this study (Figs. 1*A* and 3*A*).

If the I-A RNA binds 2′-dG by using a similar core structure adopted by guanine and adenine riboswitch aptamers, it is expected that a C-to-U mutation at the equivalent nucleotide 74 position should alter the specificity for the purine moiety of the ligand. We conducted this test by using the ligand candidate 2′-dG and its analogs, 2′-deoxyadenosine (2′-dA) and 2′-deoxy-2,6-diaminopurine nucleoside (2′-d-2,6-DAP) (Fig. 3*B*). Although 2′-dA should compensate for the aptamer C-to-U mutation, previous studies with guanine- and nucleobase 7-aminomethyl-7-deazaguanine-sensing riboswitches have revealed that 2,6-DAP binds more tightly to the mutant aptamers (27, 31). Furthermore, it has been shown that an adenine riboswitch binds 2,6-DAP with an affinity ≈30-fold better than that for adenine (27). Aptamers carrying the C-to-U mutation likely exhibit preferences for ligands that carry 2,6-DAP because of the formation of hydrogen bonds between other nucleotides in the aptamer core and the exocyclic amine at position 2 of the purine ring like those normally occurring in guanine riboswitches (Fig. 1*C*).

As expected, the unaltered I-A RNA tightly binds 2′-dG and rejects both 2′-dA and 2′-d-2,6-DAP (Fig. 3*B*). In contrast, the I-A* aptamer carrying the C-to-U mutation rejects both 2′-dG and 2′-dA (K_d values >1 mM), but binds 2′-d-2,6-DAP with a K_d of ≈8 μM. Thus, the 2,6-DAP analog also is preferred by the I-A* RNA as observed for other aptamers carrying similar C-to-U mutations. With the common form of guanine- and adenine-sensing aptamers, a U residue at the position equivalent to nucleotide 51 forms a hydrogen bond with the exocyclic amine present in guanine and 2,6-DAP (Fig. 1*C*). However, the I-A and I-A* aptamers carry a different nucleotide at the position equivalent to nucleotide 51, and therefore they likely recognize this extra amine group differently despite similarities elsewhere in the aptamer structure.

Molecular Recognition Determinants of a 2′-dG Aptamer. In addition to the various nucleoside and nucleotide analogs described earlier, we examined nucleobase analogs of guanine to further define the functional groups recognized by the I-A aptamer. Although guanine is bound by the aptamer with a poorer affinity than 2′-dG (Fig. 3*A*), guanine induces structural changes in the RNA (Fig. 4*A*) with characteristics (Fig. 4*B*) of a typical 1:1 RNA–ligand interaction (8). An important observation from the in-line probing data is that guanine does not induce the full spectrum of changes in spontaneous RNA cleavage in the J2-3 region, which are observed when 2′-dG is bound (Fig. 4*B*). This finding is consistent with the fact that the J2-3 region likely carries the nucleotides required to recognize the deoxyribose moiety of 2′-dG and, therefore, does not undergo the same level of structural stabilization with guanine that is induced by the cognate ligand. This hypothesis is further supported by the observation that guanosine, which binds with an affinity equal to that of guanine, also does not induce complete structural stabilization of the J2-3 region (Fig. 4*A*).

If the guanine base occupies the same site as does the guanine moiety of 2′-dG, then guanine analogs can be surveyed for binding activity to reveal other functional groups that are important for

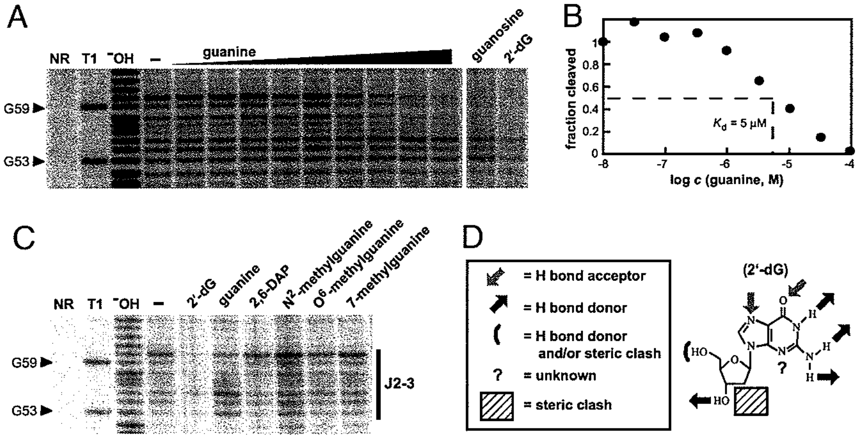


Fig. 4. Molecular recognition by a riboswitch aptamer that senses 2′-dG. (*A*) In-line probing analysis of the I-A aptamer depicted in Fig. 2*B* in guanine concentrations ranging from 0 to 100 μM or with 330 μM guanosine or 2′-dG as indicated. Other notations are as described for Fig. 2*A*. (*B*) Plot of the normalized fraction of RNA cleaved at nucleotides 59 and 60 (Fig. 2*B*) versus the logarithm of the concentration (*c*) of guanine. (*C*) In-line probing analysis of the I-A aptamer with 2′-dG, guanine, and four guanine analogs at 100 μM. Other annotations are as described for Fig. 2*A*. (*D*) Schematic representation of the molecular recognition contacts used by the 2′-dG aptamer to selectively recognize its ligand.

recognition by the aptamer. However, several guanine analogs that carry modifications of functional groups on the purine ring are not bound by the I-A aptamer when present at 100 μM (Fig. 4*C*). These findings, and those presented earlier, reveal that the I-A aptamer recognizes nearly every available functional group to form a precise binding pocket (Fig. 4*D*).

Transcription of RNAs Carrying I-A and III-B Aptamers Reveals Metabolite-Mediated Termination. The majority of guanine and adenine riboswitches are predicted to control gene expression by regulating transcription termination (J. Barrick and R.R.B., unpublished data). In these instances, the aptamer resides a short distance upstream of a predicted intrinsic transcription terminator (43, 44), which forms a strong base-paired stem followed by a run of U residues. All of the newly found aptamer variants in *M. florum* lie immediately upstream of putative intrinsic terminator stems (data not shown), indicating that they are components of riboswitches that control transcription termination.

To assess how these variant riboswitches might control gene expression, single-round transcription termination assays (45) were performed by using DNA templates containing either I-A or III-B riboswitch sequences. The amounts of terminated and full-length RNA transcripts should change in a ligand-dependent manner if the riboswitch controls this process and retains activity in an *in vitro* assay. As observed with members of several other riboswitch classes (33, 46–48), *in vitro* transcription assays measuring riboswitch control do not range the full spectrum between 100% termination and 100% full-length (FL) transcription. This result could be due to numerous differences between the reaction conditions used and the natural conditions in *M. florum* cells, such as the use of RNA polymerase from a different species, the absence of certain ions and small molecules, and the absence of proteins that might influence RNA folding. However, we do observe changes in the percentages of transcripts that are terminated when the target ligands are present in the *in vitro* transcription reaction (Fig. 5*A*). For example, the highest levels of termination for the I-A and III-B constructs occur when 2′-dG and guanine are added to the reactions, respectively. This finding matches the preferred ligands for these RNAs as determined by in-line probing assays (Fig. 3*A*).

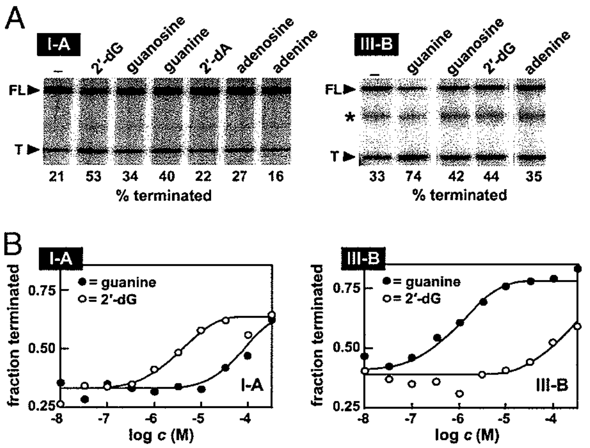


Fig. 5. *In vitro* transcription termination assays with types I-A and III-B riboswitches. (*A*) PAGE analysis of single-round transcription termination assays. Bands representing terminated transcripts (T) or full-length (FL) runoff transcripts are identified. Single-round transcription assays (see *Materials and Methods*) are conducted in the absence (–) of added ligand candidate or in the presence of 25 μM of the compounds indicated. The asterisk identifies a product of transcription pausing that appears to be unrelated to riboswitch function. (*B*) Plots of the fraction of RNAs terminated versus the logarithm of the concentration of ligand added to single-round transcription termination assays.

It has been reported that some riboswitches function as kinetically driven, rather than thermodynamically driven, gene-control elements (28, 49, 50). For example, to yield half-maximal modulation of transcription termination, kinetically driven riboswitches require a concentration of ligand much higher than the apparent K_d of the aptamer for the metabolite. The metabolite concentration needed to half-maximally modulate transcription termination, called T_{50} (49), was determined for both the I-A and III-B riboswitch constructs by examining single-round *in vitro* transcription termination assays conducted in the presence of various concentrations of 2′-dG or guanine (Fig. 5*B*). For the I-A construct, T_{50} values of 2 μM and >100 μM were observed for 2′-dG and guanine, respectively. In contrast, III-B exhibited T_{50} values for these two compounds that were nearly exactly opposite. Again, the differences in the T_{50} values for the two constructs and two ligands indicate that the type I riboswitches most likely respond to cellular 2′-dG concentrations, whereas the type III riboswitches most likely respond to guanine.

The specificities exhibited by the I-A and III-B RNAs in both assay types correspond to those predicted based on the functions of the gene products they likely control (Fig. 1*A*). The rationale for controlling expression of ribonucleotide reductase genes with 2′-dG is straightforward. In contrast, the rationale for controlling the expression of a GMP synthase gene with a guanine-sensing III-B RNA is less obvious because it might seem more logical for the riboswitch to sense the ribonucleotide product of the enzyme. However, 15 previously identified guanine riboswitches are associated with genes for GMP synthase in other organisms (ref. 26 and data not shown). Therefore, *M. florum* appears to use a variant guanine riboswitch to control this homologous gene much like other bacteria.

The discovery of variant riboswitches in *M. florum* that have altered ligand specificities and affinities suggests that a far greater diversity of riboswitches is present in bacteria and that some of these RNAs might be exceedingly narrow in phylogenetic distribution. Although highly selective for their cognate ligands, at least some riboswitch aptamers appear to be versatile and can change ligand affinities by accumulating one or only a few mutations. It seems reasonable to speculate that the common fold and consensus sequence for guanine riboswitches (Fig. 1*B*) could undergo other mutations that would allow it to bind other purine-related compounds while maintaining its secondary structure.

The two riboswitch variants that selectively sense 2′-dG have been found in only one bacterial species. Given the apparent structural versatility of RNA, there might be many unidentified riboswitch classes that reside in a single species. The aptamers for these RNAs could be close variants of known riboswitch classes, which would permit their discovery by comparative sequence analysis. However, we have evidence that structurally distinct classes of riboswitches also can be found to exist in only a few species (24). Searching for new riboswitch classes by bioinformatics methods that recognize conserved sequences or structures requires the existence of multiple copies. Therefore, the identification of exceedingly rare riboswitch classes might be best pursued by using approaches that involve direct experimental testing.

Materials and Methods

Bioinformatics. The original search for purine riboswitches was performed with the SequenceSniffer program (J. E. Barrick and R.R.B., unpublished data), with an *E*-value cutoff of 10,000. To determine whether additional purine riboswitch variants could be identified by using automated homology searches, the National Center for Biotechnology Information RefSeq database (51) was searched by using the RaveNnA extension (52) to the software package INFERNAL (www.infernal.janelia.org). A covariance model derived from all known purine riboswitches and one derived exclusively from the variant purine riboswitches were used as inputs for the searches.

Chemicals and Oligonucleotides. 2′-deoxyguanosine, 3′-deoxyguanosine, 2′-deoxyadenosine, 2′-deoxyguanosine-5′-phosphate, 2′-deoxyadenosine-5′-phosphate, 2′-deoxyguanosine-5′-diphosphate, 2′-deoxyadenosine-5′-diphosphate, 2′-deoxyguanosine-5′-triphosphate, 2′-deoxyadenosine-5′-triphosphate, guanosine-5′-phosphate, guanosine-2′-phosphate, guanosine-5′-diphosphate, guanosine-5′-triphosphate, adenosine-5′-triphosphate, 2′-deoxy-2,6-diaminopurine nucleoside (2,6-diaminopurine 2′-deoxyriboside), guanine, adenine, 2,6-diaminopurine, N^2-methylguanine, O^6-methylguanine, and 7-methylguanine were purchased from Sigma–Aldrich (St. Louis, MO). DNA oligonucleotides were synthesized by the Howard Hughes Medical Institute Keck Foundation Biotechnology Resource Center at Yale University; purified by denaturing PAGE; eluted from the gel by crush-soaking in 10 mM Tris·HCl (pH 7.5 at 23°C), 200 mM NaCl, and 1 mM EDTA; and precipitated with ethanol.

In-Line Probing Assays. RNA constructs were prepared from synthetic double-stranded DNA templates by *in vitro* transcription by using methods similar to those described previously (53). The resulting RNAs were dephosphorylated by using alkaline phosphatase (Roche Diagnostics, Indianapolis, IN) and subsequently labeled with ^{32}P by using T4 polynucleotide kinase (New England Biolabs, Ipswich, MA) following the manufacturer's instructions. Radiolabeled RNAs ($\approx$2 nM) were subjected to in-line probing by incubation with or without various ligands for 40 h in 10-μl reactions containing 50 mM Tris·HCl (pH 8.5 at 23°C), 20 mM $MgCl_2$, and 100 mM KCl. Denaturing 10% PAGE was used to separate spontaneously cleaved products, which were visualized by using a Molecular Dynamics PhosphorImager (Sunnyvale, CA). ImageQuaNT software was used to quantitate spontaneous cleavage amounts.

***In Vitro* Transcription Termination Assays.** The protocol for single-round transcription assays was adapted from that described previously (45). The *lysC* promoter of *B. subtilis* was used to facilitate greater transcription yield with the I-A construct. Transcription reactions contained 100 nM DNA template in 20 mM Tris·HCl (pH 8.0 at 23°C)/20 mM NaCl/14 mM $MgCl_2$/0.1 mM EDTA/1 mg/ml BSA/50% glycerol 2.2 μM *E. coli* RNA polymerase holoenzyme (Epicenter Technologies, Madison, WI). Transcription was initiated by adding 2.5 μM GTP and UTP, 1 μM ATP, 4 μCi [α-^{32}P]ATP, and 1.35 μM ApA dinucleotide. After incubating for 10 min at 37°C, 0.075 mM GTP, ATP, and CTP; 0.025 mM UTP; and 0.1 mg/ml heparin were added, and the resulting mixture was allowed to incubate for 20 min at 37°C. Products were separated by denaturing 6% PAGE and imaged and quantitated by using a PhosphorImager and ImageQuaNT software.

The FL and T transcript amounts were established by correcting for the differences in the number of A residues in the molecules. The percentage of [α-^{32}P]ATP compared with total ATP concentration in the initiation and elongation reactions (7% and 0.4%, respectively) was established, and the relative amount of radioactivity per T (U_T) and FL (U_{FL}) transcripts was calculated for each transcript size by using the following equation: [(Number of A residues in initiation region)(7%) + (Number of A residues in elongation region)(0.4%)] = U.

U_T/U_{FL} is equal to the correction factor (X%) that accounts for the increased number of radiolabeled adenosine residues in the FL transcript. The equation used to establish the percentage of transcription termination was: [T/(T + FL)(X%)] = percentage termination.

We thank Beth Grove for performing automated homology searches, Dr. Tom Knight (Massachusetts Institute of Technology, Cambridge, MA) for providing *Mesoplasma florum* genomic DNA, members of the R.R.B. laboratory for helpful discussions, and Dr. Narasimhan Sudarsan for comments on experiments. This work was supported by the Howard Hughes Medical Institute, National Institutes of Health Grants GM068819 and DK070270, and a National Science Foundation predoctoral fellowship (to J.N.K.).

1. Nahvi A, Sudarsan N, Ebert MS, Zou X, Brown KL, Breaker RR (2002) *Chem Biol* 9:1043–1049.
2. Mandal M, Breaker RR (2004) *Nat Rev Mol Cell Biol* 5:451–463.
3. Soukup JK, Soukup GA (2004) *Curr Opin Struct Biol* 14:344–349.
4. Winkler WC (2005) *Curr Opin Chem Biol* 9:594–602.
5. Winkler WC, Breaker RR (2005) *Annu Rev Microbiol* 59:487–517.
6. Mandal M, Lee M, Barrick JE, Weinberg Z, Emilsson GM, Ruzzo WL, Breaker RR (2004) *Science* 306:275–279.
7. Sudarsan N, Hammond MC, Block KF, Welz R, Barrick JE, Roth A, Breaker RR (2006) *Science* 314:300–304.
8. Welz R, Breaker RR (2007) *RNA* 13:573–582.
9. Stoddard CD, Batey RT (2006) *ACS Chem Biol* 1:751–754.
10. Batey RT, Gilbert SD, Montange RK (2004) *Nature* 432:411–415.
11. Serganov A, Yuan YR, Pikovskaya O, Polonskaia A, Malinina L, Phan AT, Hobartner C, Micura R, Breaker RR, Patel DJ (2004) *Chem Biol* 11:1729–1741.
12. Thore S, Leibundgut M, Ban N (2006) *Science* 312:1208–1211.
13. Serganov A, Polonskaia A, Phan AT, Breaker RR, Patel DJ (2006) *Nature* 441:1167–1171.
14. Montange RK, Batey RT (2006) *Nature* 441:1172–1175.
15. Kline DJ, Ferré-D'Amaré AR (2006) *Science* 313:1752–1756.
16. Edwards TE, Ferré-D'Amaré AR (2006) *Structure (London)* 14:1459–1468.
17. Cochrane J, Lipchock S, Strobel S (2007) *Chem Biol* 14:97–105.
18. Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS (2002) *J Biol Chem* 277:48949–48959.
19. Barrick JE, Corbino KA, Winkler WC, Nahvi A, Mandal M, Collins J, Lee M, Roth A, Sudarsan N, Jona I, *et al.* (2004) *Proc Natl Acad Sci USA* 101:6421–6426.
20. Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS (2003) *Nucleic Acids Res* 31:6748–6757.
21. Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS (2003) *J Biol Chem* 278:41148–41159.
22. Nahvi A, Barrick JE, Breaker RR (2004) *Nucleic Acids Res* 32:143–150.
23. Corbino KA, Barrick JE, Lim J, Welz R, Tucker BJ, Puskarz I, Mandal M, Rudnick ND, Breaker RR (2005) *Genome Biol* 6:R70.
24. Weinberg Z, Barrick JE, Yao Z, Roth A, Kim JN, Gore J, Wang JX, Lee ER, Block KF, Sudarsan N, *et al.* (2007) *Nucleic Acids Res* 4809–4819.
25. Breaker RR (2006) In *The RNA World*, eds Gesteland RF, Cech TR, Atkins JF (Cold Spring Harbor Lab Press, Cold Spring Harbor, NY), 3rd Ed, pp 89–107.
26. Mandal M, Breaker RR (2003) *Cell* 113:577–586.
27. Mandal M, Breaker RR (2004) *Nat Struct Mol Biol* 11:29–35.
28. Wickiser JK, Cheah MT, Breaker RR, Crothers DM (2005) *Biochemistry* 44:13404–13414.
29. Lemay J-F, Penedo JC, Tremblay R, Lilley DMJ, Lafontaine DA (2006) *Chem Biol* 13:857–868.
30. Rieder R, Lang K, Graber D, Micrua R (2007) *ChemBioChem* 8:896–902.
31. Roth A, Winkler WC, Regulski EE, Lim J, Jona I, Barrick JE, Ritwik A, Kim J, Iwata-Reuyl D, Breaker RR (2007) *Nat Struct Mol Biol* 14:308–317.
32. Fuchs RT, Grundy FJ, Henkin TM (2006) *Nat Struct Mol Biol* 13:226–233.
33. McDaniel BA, Grundy FJ, Artsimovitch I, Henkin TM (2003) *Proc Natl Acad Sci USA* 100:3083–3088.
34. Winkler WC, Nahvi A, Sudarsan N, Barrick JE, Breaker RR (2003) *Nat Struct Biol* 10:701–707.
35. Epshtein V, Mironov AS, Nudler E (2003) *Proc Natl Acad Sci USA* 100:5052–5056.
36. Hutchison CA III, Montague MG (2002) In *Molecular Biology and Pathogenicity of Mycoplasmas*, eds Razin S, Herrmann R (Kluwer Academic/Plenum, New York), pp 221–253.
37. Jordan A, Reichard P (1998) *Annu Rev Biochem* 67:71–98.
38. Nordlund P, Reichard P (2006) *Annu Rev Biochem* 75:681–706.
39. Soukup GA, Breaker RR (1999) *RNA* 5:1308–1325.
40. Noeske J, Richter C, Grundl MA, Nasiri HR, Schwalbe H, Wohnert J (2005) *Proc Natl Acad Sci USA* 102:1372–1377.
41. Bengert P, Dandekar T (2004) *Nucleic Acids Res* 32:W154–W159.
42. Lemay J-F, Lafontaine DA (2007) *RNA* 13:339–350.
43. Gusarov I, Nudler E (1999) *Mol Cell* 3:495–504.
44. Yarnell WS, Roberts JW (1999) *Science* 284:611–615.
45. Landick R, Wang, D, Chan CL (1996) *Methods Enzymol* 274:334–353.
46. Sudarsan N, Wickiser JK, Nakamura S, Ebert MS, Breaker RR (2003) *Genes Dev* 17:2688–2697.
47. Winkler WC, Nahvi A, Sudarsan N, Barrick JE, Breaker RR (2003) *Nat Struct Biol* 10:701–707.
48. Blount KF, Wang JX, Lim J, Sudarsan N, Breaker RR (2007) *Nat Chem Biol* 3:44–49.
49. Wickiser JK, Winkler WC, Breaker RR, Crothers DM (2005) *Mol Cell* 18:49–60.
50. Gilbert SD, Stoddard CD, Wise SJ, Batey RT (2006) *J Mol Biol* 359:754–768.
51. Pruitt KD, Tatusova T, Maglott DR (2005) *Nucleic Acids Res* 35:D61–D65.
52. Weinberg Z, Ruzzo WL (2006) *Bioinformatics* 22:35–39.
53. Roth A, Nahvi A, Lee M, Jona I, Breaker RR (2006) *RNA* 12:607–619.
54. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV (2001) *Nucleic Acids Res* 29:22–28.

nature
structural &
molecular biology

Adenine riboswitches and gene activation by disruption of a transcription terminator

Maumita Mandal & Ronald R Breaker

A class of riboswitches that recognizes guanine and discriminates against other purine analogs was recently identified. RNAs that carry the consensus sequence and structural features of guanine riboswitches are located in the 5′ untranslated region (UTR) of numerous prokaryotic genes, where they control the expression of proteins involved in purine salvage and biosynthesis. We report that three representatives of this riboswitch class bind adenine with values for apparent dissociation constant (apparent K_d) that are several orders of magnitude lower than those for binding guanine. Because preference for adenine is attributable to a single nucleotide substitution, the RNA most likely recognizes its ligand by forming a Watson-Crick base pair. In addition, the adenine riboswitch associated with the *ydhL* gene of *Bacillus subtilis* functions as a genetic 'on' switch, wherein adenine binding causes a structural rearrangement that precludes formation of an intrinsic transcription terminator stem.

Guanine-sensing riboswitches are RNA genetic control elements that modulate gene expression in response to changing concentrations of guanine[1]. They represent one of several classes of metabolite-binding riboswitches that regulate gene expression in response to various fundamental compounds such as lysine and the coenzymes FMN, SAM, B_{12} and TPP (thiamine pyrophosphate)[1,2]. Typically, each riboswitch comprises two functional domains, an aptamer and an expression platform, that function together to transduce chemical signals into altered patterns of gene expression. The aptamer is a specific receptor for the target metabolite, and ligand binding causes allosteric changes in both the aptamer and expression platform domains.

We have closely examined the ligand specificities of the natural aptamers from guanine- and lysine-specific riboswitches[1,3], and less comprehensively examined the FMN, SAM, B_{12} and TPP aptamers[4–7]. In each case, the RNAs closely discriminate among different molecules, disfavoring the binding of even closely related metabolite analogs. This discrimination is a hallmark of enzymes and receptors, including genetic regulatory factors, that carry out precise biological processes in the presence of complex chemical mixtures.

The molecular recognition characteristics of guanine riboswitches are distinguished by the fact that nearly every position around the purine heterocycle seems critical for high-affinity binding by the aptamer. Thus, the arrangement of the binding pocket permits the riboswitch to control gene expression in response to changing guanine concentrations, but prevents modulation of gene expression in response to increasing concentrations of adenine[1,8]. However, receptors made of RNA, like their protein counterparts, could acquire altered molecular recognition characteristics as a result of natural selection. This would allow riboswitches to emerge that selectively sense and respond to proximal metabolites in metabolic pathways.

Here, we present evidence of a variant class of riboswitches that responds to adenine. These riboswitches carry an aptamer domain similar in sequence and secondary structure to the guanine aptamer described recently[1]. However, each representative of the adenine subclass of riboswitches carries a C→U mutation in the conserved core of the aptamer, suggesting that this residue is involved in metabolite recognition. Our findings indicate that the identity of this single nucleotide determines the binding specificity between guanine and adenine, providing an example of how complex riboswitch structures could mutate to recognize new metabolite targets.

RESULTS

Phylogenetic comparison between riboswitch domains

In a recent study, we used a comparative sequence strategy to identify a series of intergenic regions from several prokaryotic species that carry a conserved sequence element called the 'G box'[1]. *B. subtilis* carries at least five of these motifs, which were also identified using genetic techniques[9]. Each representative of the phylogeny has three potential base-paired elements (P1–P3) and as many as 24 nucleotides that are conserved in >90% of the examples identified so far. We present here a subset of this phylogeny with features common to the G box motif highlighted (Fig. 1a). Selected representatives seem to be expressed as part of the mRNA transcript of the gene immediately downstream, and thus are present as RNA elements in the 5′ UTR of certain mRNAs.

We identified several notable differences in the guanine-binding domain of *xpt* (Fig. 1b) relative to the RNA from *ydhL* (Fig. 1c), which encodes a putative purine efflux pump[9]. First, among the 23 sequence variations in *ydhL* as compared with *xpt*, 20 reside in base-paired elements and base pairing is retained with most of these changes. This strongly indicates that the overall secondary structure of the two RNAs

Department of Molecular, Cellular and Developmental Biology, Yale University, PO Box 208103, New Haven, Connecticut 06520-8103, USA. Correspondence should be addressed to R.R.B. (ronald.breaker@yale.edu).

Published online 29 December 2003; doi:10.1038/nsmb710

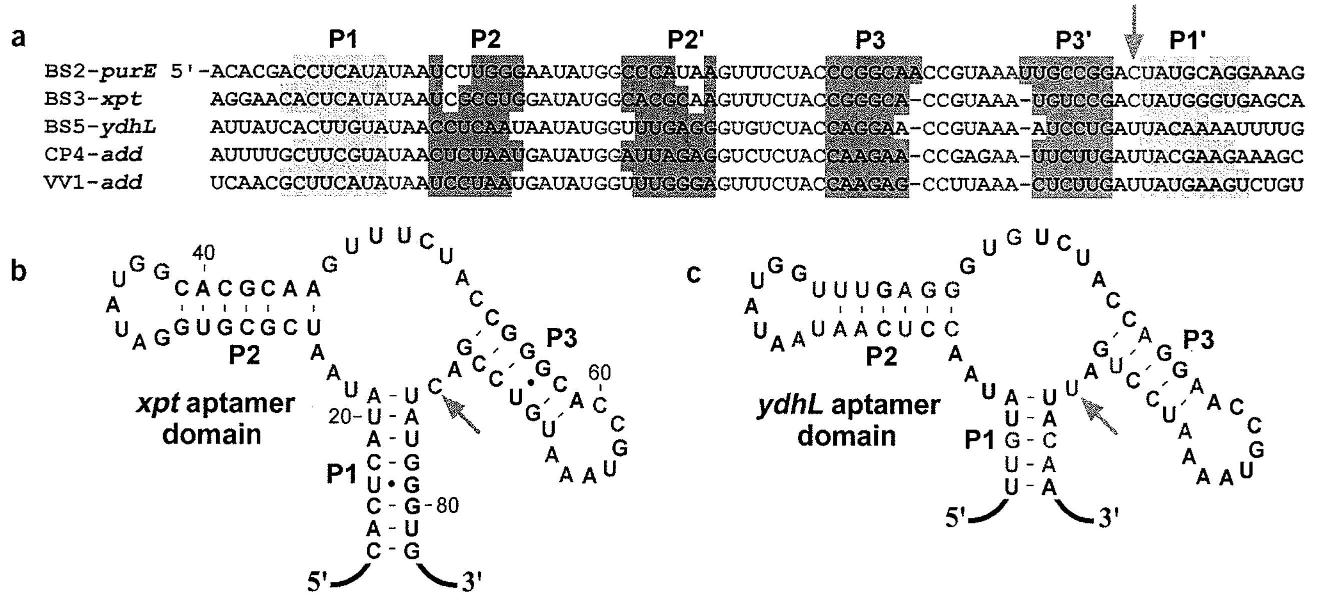

a
P1
P2
P2'
P3
P3'
P1'
BS2-purE 5'-ACACGACCUCAUAUAAUCUUGGGAAUAUGGCCCAUAAGUUUCUACCCGGCAACCGUAAAUUGCCGGACUAUGCAGGAAAG
BS3-xpt AGGAACACUCAUAUAAUCGCGUGGAUAUGGCACGCAAGUUUCUACCGGGCA-CCGUAAA-UGUCCGACUAUGGGUGAGCA
BS5-ydhL AUUAUCACUUGUAUAACCUCAAUAAUAUGGUUUGAGGGUGUCUACCAGGAA-CCGUAAA-AUCCUGAUUACAAAAUUUUG
CP4-add AUUUUGCUUCGUAUAACUCUAAUGAUAUGGAUUAGAGGUCUCUACCAAGAA-CCGAGAA-UUCUUGAUUACGAAGAAAGC
VV1-add UCAACGCUUCAUAUAAUCCUAAUGAUAUGGUUUGGGAGUUUCUACCAAGAG-CCUUAAA-CUCUUGAUUAUGAAGUCUGU
b
xpt aptamer domain
P1
P2
P3
20
40
60
80
5'
3'
c
ydhL aptamer domain
P1
P2
P3
5'
3'

Figure 1 Guanine- and adenine-specific riboswitches. (a) Sequence and structural features of the two guanine-specific (*purE* and *xpt*) and three adenine-specific aptamer domains examined in this study. P1–P3, three base-paired stems composing the secondary structure of the aptamer domain. Red letters, positions whose base identity is conserved in >90% of representatives in the phylogeny[1]. Arrow, a nucleotide in the conserved core of the aptamer that determines ligand specificity. BS, CP and VV designate *B. subtilis, Clostridium perfringens* and *Vibrio vulnificus*, respectively. (b,c) Sequence and secondary structure of the *xpt* (b) and *ydhL* (c) aptamers. Green letters, positions in the *ydhL* aptamer that differ from those in the *xpt* aptamer. Nucleotides in *xpt* are numbered as described[1]. Other notations are as described in a.

is similar. Second, the remaining three mutations are in unpaired regions, such that two (corresponding to positions 32 and 48 relative to *xpt*) reside at locations known to be variable. These mutations probably do not substantially impact the structure and function of the RNA. Third, the remaining mutation is 74C→U (position relative to *xpt*), which otherwise corresponds to a strictly conserved nucleotide of the three-stem junction. Given the location of this mutation, we suspected that this change might alter the molecular recognition characteristics of the *ydhL* aptamer.

Variant G box RNAs selectively bind adenine

The *xpt* aptamer makes numerous contacts with its ligand, and as many as seven hydrogen bonds might be involved in forming the RNA–ligand complex[1]. Furthermore, there is evidence that steric clashes probably aid in restricting the range of metabolites that can be bound by the RNA. This array of contacts can be established only by the formation of multiple interactions between guanine and distal parts of the RNA.

An intriguing possibility is that C74 of *xpt* could form a Watson-Crick base pair with guanine, thus forming three of these hydrogen bonds. Because a C→U mutation resides in the corresponding position in *B. subtilis ydhL* and two RNAs from *Clostridium perfringens* and *Vibrio vulnificus*, we speculated that these RNAs might be adenine-responsive riboswitches. The latter two genes (*add*) encode adenine deaminase enzymes, lending further support to this hypothesis. It seems reasonable that the concentration of adenine would be monitored to determine the expression levels of adenine deaminase.

We examined the ligand specificity of five G box RNAs (Fig. 1a) using in-line probing[10,11] in which we monitored the spontaneous cleavage of RNA in the absence of ligand, or in the presence of guanine or adenine. As predicted previously[1], the *purE* RNA (Fig. 2a) exhibits changes in the pattern of spontaneous cleavage products in the presence of guanine that correspond to those observed for the *xpt* RNA (Fig. 2b). These results confirm that the *purE* RNA, like the *xpt* RNA, responds allosterically to guanine and not to adenine when incubated in the presence of the concentrations of ligand tested.

In contrast, all three RNAs that carry the C→U mutation in the junction between P1 and P3 (corresponding to C74 of *xpt*) do not respond to guanine, but show structural modulation only when incubated in the presence of adenine. Furthermore, the patterns of spontaneous cleavage for the adenine-specific aptamers are consistent with the secondary-structure model proposed for G box RNAs (Fig. 1). These results indicate that certain variants of the G box class of RNAs are adenine sensors. They also suggest that, in their natural settings, the *ydhL* RNA from *B. subtilis* and the two *add* RNAs from *C. perfringens* and *V. vulnificus* are adenine-specific riboswitches.

ydhL RNA binds adenine with high affinity and selectivity

The aptamer domains of riboswitches bind their corresponding target compounds tightly, and they discriminate against analogs, in some cases by orders of magnitude in apparent K_d. For example, the guanine riboswitch from *B. subtilis xpt* has an apparent K_d for guanine of ~5 nM, but binds adenine with an apparent K_d value that is at least 100,000-fold higher. We have used in-line probing assays to determine the binding affinities of an 80-nucleotide portion of the *B. subtilis ydhL* RNA (80 *ydhL*) for these two purines. As expected, the RNA exhibits progressively changing patterns of spontaneous RNA cleavage fragments as adenine concentration increases (Fig. 3a), but the pattern remains unchanged as guanine concentration increases as high as 10 μM (see below).

We grouped the bands corresponding to spontaneous cleavage fragments that undergo changes as adenine concentration increases into four sites and quantified the extent of cleavage relative to the total RNA present. We plotted these data (Fig. 3b) to estimate the apparent K_d for

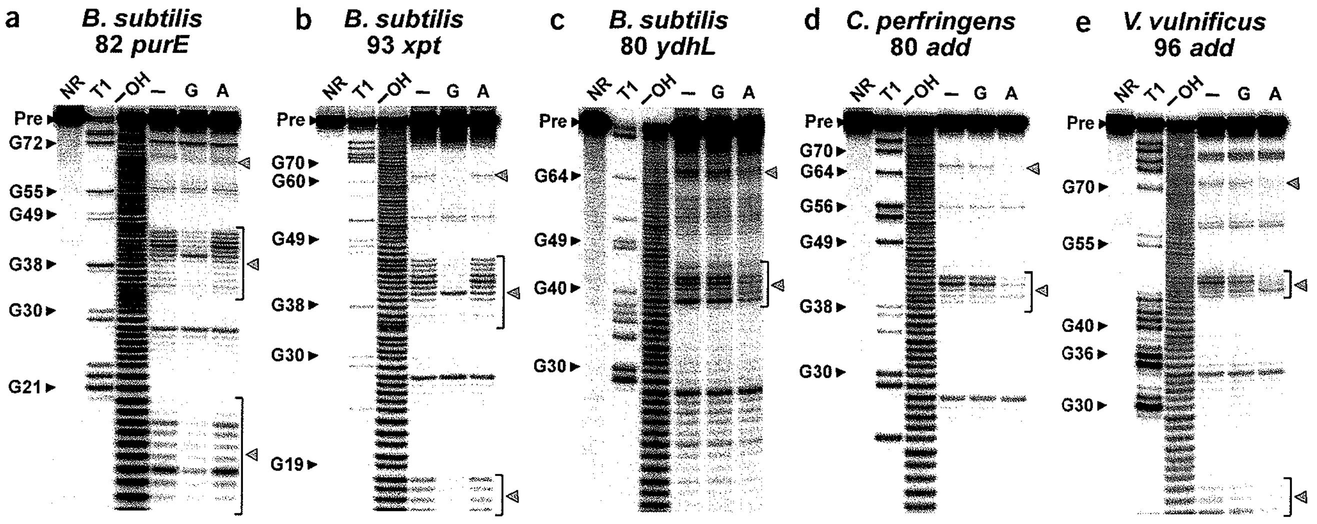


Figure 2 Ligand specificity of five G box RNAs. (a–e) In-line probing assays for the conserved aptamer domains as labeled. NR, T1 and ⁻OH: marker lanes in which precursor RNAs (Pre) were not incubated, or were partially digested with RNase T1 or alkali, respectively. Selected bands corresponding to RNase T1 digestion (cleavage 3′ relative to guanidyl residues) are labeled for each RNA. RNAs were incubated for 40 h in the absence of ligand (–), or in the presence of 1 μM guanine (G) or adenine (A). Large arrowheads, sites of substantial change in cleavage pattern due to the addition of a particular ligand. See Methods for additional details.

ligand binding. In this instance, a half-maximal decrease in spontaneous cleavage at sites 1, 2 and 4, and a corresponding half-maximal increase in spontaneous cleavage at site 3 occurs with ~300 nM adenine in the in-line probing assay. Thus, the *ydhL* aptamer binds adenine with an apparent K_d similar to those of other classes of riboswitches.

We further examined the molecular recognition characteristics of 80 *ydhL* using the same in-line probing strategy with a variety of analogs. For example, a series of purine analogs that are close chemical variants of adenine, 2,6-DAP, 2-AP, P and MA, bind measurably to the RNA (ligands listed in order of decreasing affinity) (Fig. 4a). Furthermore, the relative affinities of the RNA for various ligands indicate contact points that the aptamer may use to establish molecular recognition (Fig. 4a, bottom right). This model is consistent with our findings that a series of purine analogs do not bind measurably to the 80 *ydhL* RNA (Fig. 4b).

The collection of purines recognized by 80 *ydhL* indicate that only the Watson-Crick base-pairing face of the purine ligand is recognized differently by the *ydhL* aptamer as compared with the *xpt* aptamer. For example, modification at the C8 position (8-chloroadenine) prevents ligand binding, suggesting that a steric clash between certain purines and 80 *ydhL* occurs, as was observed for the *xpt* aptamer[1]. Notably, 2,6-DAP, and not adenine, is the tightest-binding ligand for 80 *ydhL*; this provides further insight into the similarities between the *ydhL* and *xpt* aptamers. The *xpt* aptamer has a poorer binding affinity for N^2-methylguanine than for both guanine and hypoxanthine, indicating that the RNA might make use of two hydrogen bond contacts with the 2-amino group[1]. One of these could be explained by the formation of a G-C base pair, whereas the other interaction with this functional group of the ligand would need to be formed by a different part of the aptamer. Therefore the 80 *ydhL* RNA would lose the ability to form one hydrogen bond to the 2-amino group that otherwise would be formed by the 6-keto group of cytosine, but retain the other because this part of the aptamer remains the same between the two RNAs. Thus, the molecular recognition characteristics of these RNAs are consistent with the hypothesis that the *ydhL* RNA differs in molecular recognition from *xpt* with a pattern that can be explained by a change from a Watson-Crick G-C base pair in *xpt* to a Watson-Crick A-U base pair in *ydhL*.

Swapping ligand specificity of G box RNAs

We used molecular engineering to test the hypothesis that the *xpt* and *ydhL* RNAs derive their specificity for guanine or adenine by a Watson-Crick base-pairing interaction. A similar approach was used previously[12] to change the ligand-rescue specificity of an abasic hammerhead ribozyme construct from guanine to adenine. Both wild-type (93 *xpt* and 80 *ydhL*) and mutant (93 *xpt* 74C→U and 80 *ydhL* 74U→C) forms of G box aptamers were generated and tested for binding activity with guanine and adenine (Fig. 5).

As observed previously[1], the aptamer based on *xpt* undergoes structural modulation only when incubated in the presence of guanine, and shifts the distribution of tritiated guanine (but not adenine) in an equilibrium dialysis assay (Fig. 5a). However, the 93 *xpt* RNA that carries a single 74C→U mutation is unresponsive to guanine, but exhibits structural modulation and binding activity during equilibrium dialysis only in the presence of adenine (Fig. 5b). Similarly, the K_d of the wild-type *xpt* RNA for 2,6-DAP is 10 μM, whereas the C→U mutant version of this RNA binds 2,6-DAP with a K_d of ~10 nM (data not shown). In contrast, the wild-type 80 *ydhL* RNA is specific for adenine (Fig. 5c), whereas the corresponding U→C mutation alters binding specificity to guanine (Fig. 5d). Therefore, we conclude that the primary determinant of the base specificity of G box aptamers is the cytosine or uracil residue that is present in the junction between stems P1 and P3, and that this base most likely forms a conventional Watson-Crick base pair with its target ligand.

Genetic control mechanism of the *ydhL* RNA

In most instances, riboswitches control gene expression in prokaryotes by allosteric interconversion between alternate base-paired structures[2]. For example, a TPP riboswitch from the *thiM* gene of *Escherichia coli*

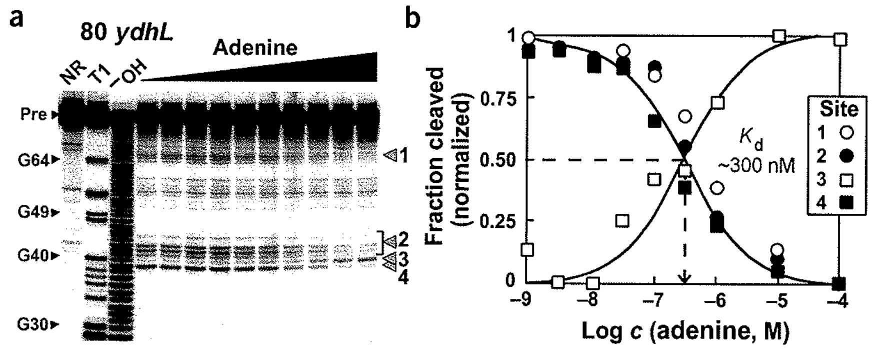


Figure 3 Binding affinity of the *ydhL* aptamer for adenine. (**a**) In-line probing assay for the 80 *ydhL* RNA at various concentrations of adenine. For each lane, sites 1–4 were quantified and the fraction of RNA cleaved was used to determine the apparent K_d. (**b**) Plot of the normalized fraction of RNA that has undergone spontaneous cleavage at sites 1–4 versus the concentration of adenine. See Methods for additional details.

makes use of alternate base pairing to sequester the Shine-Dalgarno sequence of the mRNA in the presence of ligand; this presumably results in reduced translation initiation[7]. In contrast, TPP riboswitches from *B. subtilis* harness ligand-binding events to alter base-pairing patterns and form intrinsic terminator stems that cause transcription elongation to abort[13,14]. Similarly, metabolite-mediated formation of transcription terminator stems is used by certain riboswitches that respond to FMN[4,15], SAM[5,16,17], guanine[1] and lysine[3].

We examined the UTR sequence of the *ydhL* riboswitch for evidence of a transcription termination mechanism. Consistent with this possibility, the 5′-UTR of the *ydhL* mRNA can form a large hairpin, composed of as many as 22 base pairs, followed by a run of eight uridyl residues (Fig. 6a). This structural feature, which was also noted elsewhere recently[9], is characteristic of an exceptionally long intrinsic terminator stem. We speculate that in the absence of adenine the riboswitch forms this intrinsic terminator. In this case, the default genetic control status of this riboswitch would be the predicted 'off' state, which prevents gene expression by inducing transcription termination. In the presence of adenine, gene expression would be expected to proceed because a substantial portion of the left shoulder of the terminator stem would be required to form stems P1 and P3 of the adenine aptamer domain. Because stems P1 and P2 are integral components of the adenine aptamer, ligand binding would establish a structure that precludes formation of the terminator stem.

We assessed this proposed mechanism for the *ydhL* riboswitch *in vivo* by generating reporter constructs in which various forms of guanine- and adenine-specific riboswitches were integrated into the *B. subtilis* genome. As controls, we prepared two reporter constructs with either the wild-type *xpt* riboswitch or the *xpt* variant with the 74C→U mutation. As expected, the wild-type *xpt* construct caused repression of β-galactosidase expression with excess guanine in the culture medium (Fig. 6b). This finding is similar to those reported previously for the function of the guanine riboswitch from *xpt*[1]. Adenine also shows moderate (about four-fold) repression of reporter expression after a 6 h incubation. This latter effect is most likely attributable to the PurR protein, which moderately downregulates transcription initiation in response to adenine at the *xpt-pbuX* promoter used in this construct[8].

A nearly identical *xpt* construct carrying the C→U mutation causes a loss of regulation upon addition of guanine, but shows no change in the putative protein-dependent control due to adenine (Fig. 6c). These results are consistent with the observed loss of guanine binding *in vitro* with this mutation, but suggest that the resulting specificity change to adenine *in vitro* does not permit robust adenine-dependent genetic control *in vivo*. The diminished expres-

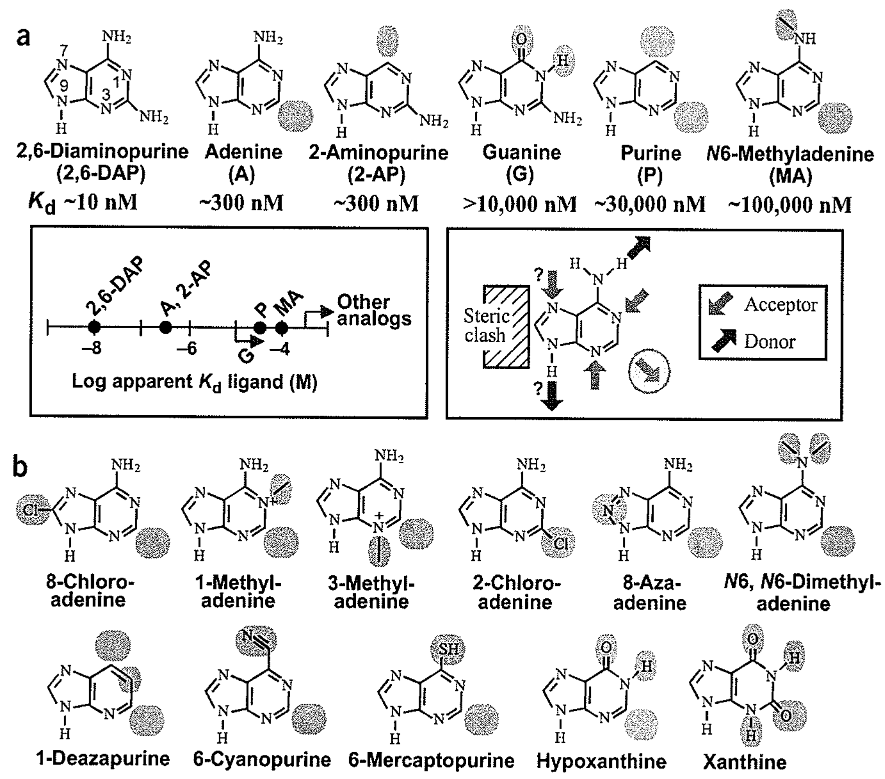


Figure 4 Specificity of molecular recognition by the adenine aptamer from *ydhL*. (**a**) Top: Chemical structures of adenine, guanine and other purine analogs that bind measurably to the 80 *ydhL* RNA. Chemical changes relative to 2,6-DAP, the tightest-binding compound, are highlighted in pink. Bottom left: plot of apparent K_d values for various purines. Bottom right: model for the chemical features of adenine that are molecular recognition contacts for *ydhL*. The importance of N7 and N9 has not been determined. Encircled arrow indicates that a contact could exist if a hydrogen bond donor were appended to C2. (**b**) Chemical structures of various purines that are not bound by the 80 *ydhL* RNA (K_d values >300 μM).

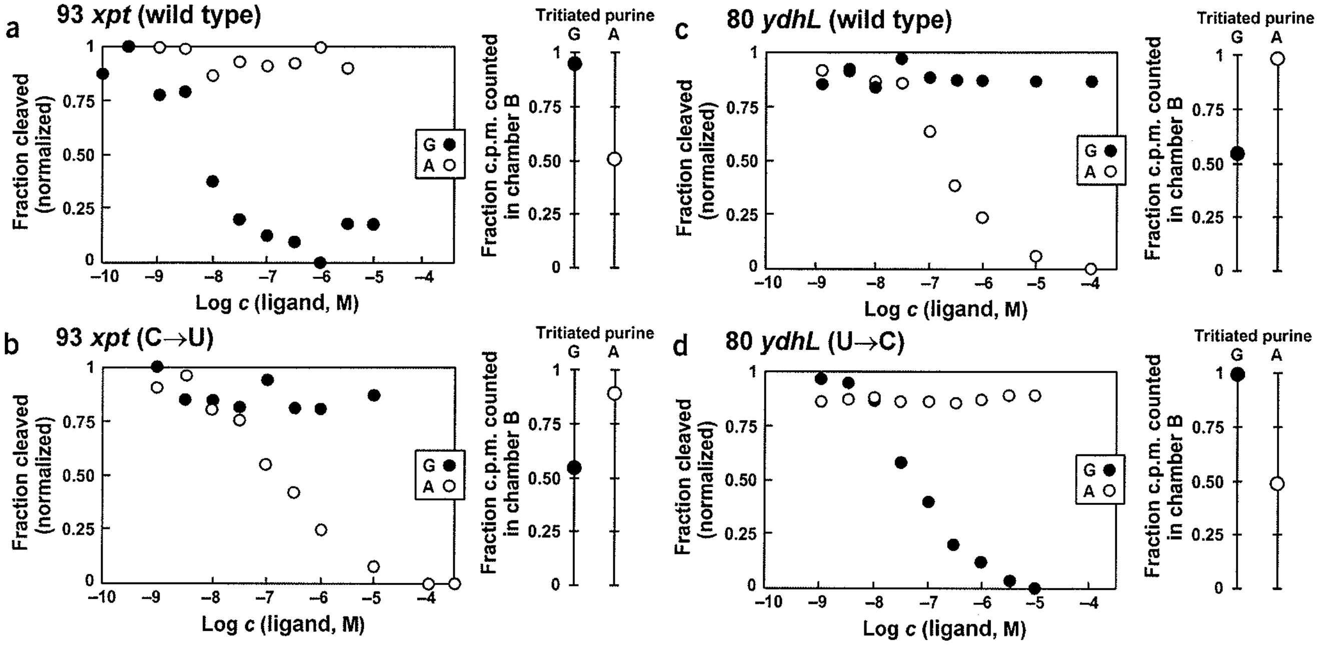


Figure 5 Interconversion of guanine- and adenine-specific aptamers. (**a**) Left: plot of the normalized fraction of wild-type 93 *xpt* RNA cleavage product for a given site versus the logarithm of the concentration of ligand present during incubation in an in-line probing assay. Cleavage products monitored for modulation correspond to site 3 (**Fig. 3a**). Right: plot of the fraction of the total counts per minute (c.p.m.) in chamber B relative to the total c.p.m. from sides A and B of an equilibrium dialysis chamber. Values of ~0.5 indicate an equal distribution of ligand (no binding) whereas values of ~1 indicate that most of the ligand is bound to the RNA in side B of the chamber. (**b–d**) In-line probing plots and equilibrium dialysis plots for 93 *xpt* (C→U mutation), 80 *ydhL* and 80 *ydhL* (U→C mutation), respectively. Details are described in **a** or in Methods.

sion upon addition of adenine is again probably due to the PurR protein.

In contrast to the *xpt* riboswitch, the operation of the corresponding wild-type and mutant *ydhL* reporter constructs indicates that the latter is an adenine-dependent riboswitch with the opposite response to rising levels of ligand. Specifically, the wild-type *ydhL* construct has very low β-galactosidase activity when assayed in the absence of ligand, or in the presence of guanine (Fig. 6d). However, a more than ten-fold increase in gene expression occurs in response to added adenine. In addition, the single U→C mutation in the P1-P3 junction of the aptamer causes substantial (~100-fold) derepression regardless of ligand (Fig. 6e). Although this seems to contradict the model proposed for *ydhL* riboswitch function, this mutation indeed disrupts adenine binding, but it also causes a mismatch in the terminator stem. If this mismatch sufficiently destabilizes the terminator stem, or adversely affects the folding pathway for the riboswitch, then the default 'off' status for the genetic control element would be expected to change to a default 'on' status. Therefore, the observed level of gene expression might indicate full activation of the *ydhL* gene when its genetic control element is unaffected by the concentrations of purines in the cell.

DISCUSSION

Structure and evolution of adenine riboswitches

The sequence and biochemical similarities between guanine- and adenine-specific G box RNAs indicate that they are analogous in overall secondary and tertiary structure. The ease of interchanging ligand specificities of these aptamers through single mutations to the *xpt* and *ydhL* aptamers suggests that such changes might occur with high frequency in natural populations. However, neither single-base variant of the *xpt* or *ydhL* riboswitches exhibits corresponding specificity changes in genetic control *in vivo*, suggesting that multiple mutations might be necessary to make a useful change in riboswitch specificity.

The binding affinity of the resulting single-base *xpt* variant is not as robust for its new ligand. Specifically, the wild-type *xpt* RNA has an apparent K_d for guanine of ≤5 nM (Fig. 5a), whereas the C→U variant of this RNA has an apparent K_d for adenine of ~100 nM (Fig. 5b). In this case, although the mutation causes a substantial change in base discrimination between guanine and adenine, binding affinity for the matched ligand has been somewhat degraded. In contrast, the wild-type and mutant *ydhL* RNAs exhibit both specificity change and retention of binding affinity for the matched ligands (Fig. 5c,d). However, the affinity for the U→C variant of 80 *ydhL* for guanine seems at least ten-fold lower than that of 93 *xpt*.

These observations, along with the operation of the variant RNAs *in vivo*, suggest a complex issue regarding riboswitch function inside cells. Current studies are more comprehensively addressing the issue of ligand concentration and genetic control. Certain riboswitches seem to be kinetically driven such that genetic control is determined by the rate constant for ligand association and not by the thermodynamic parameter K_d (unpublished data). Although the adenine-binding riboswitch engineered by mutating a single base of the *xpt* RNA has a K_d similar to that of the natural adenine riboswitch from *ydhL*, other mutations might be required to reestablish the kinetic parameters needed for function *in vivo*.

Thus, accessory mutations that do not directly define ligand specificity but that further adjust binding affinity might be necessary for G box RNAs to interconvert between guanine and adenine ligands in a biological setting. In this regard, it is notable that the *ydhL* and *xpt* aptamers differ from each other at 23 positions (Fig. 1), only one of which is an obviously critical position (C74 of *xpt*). Some of these

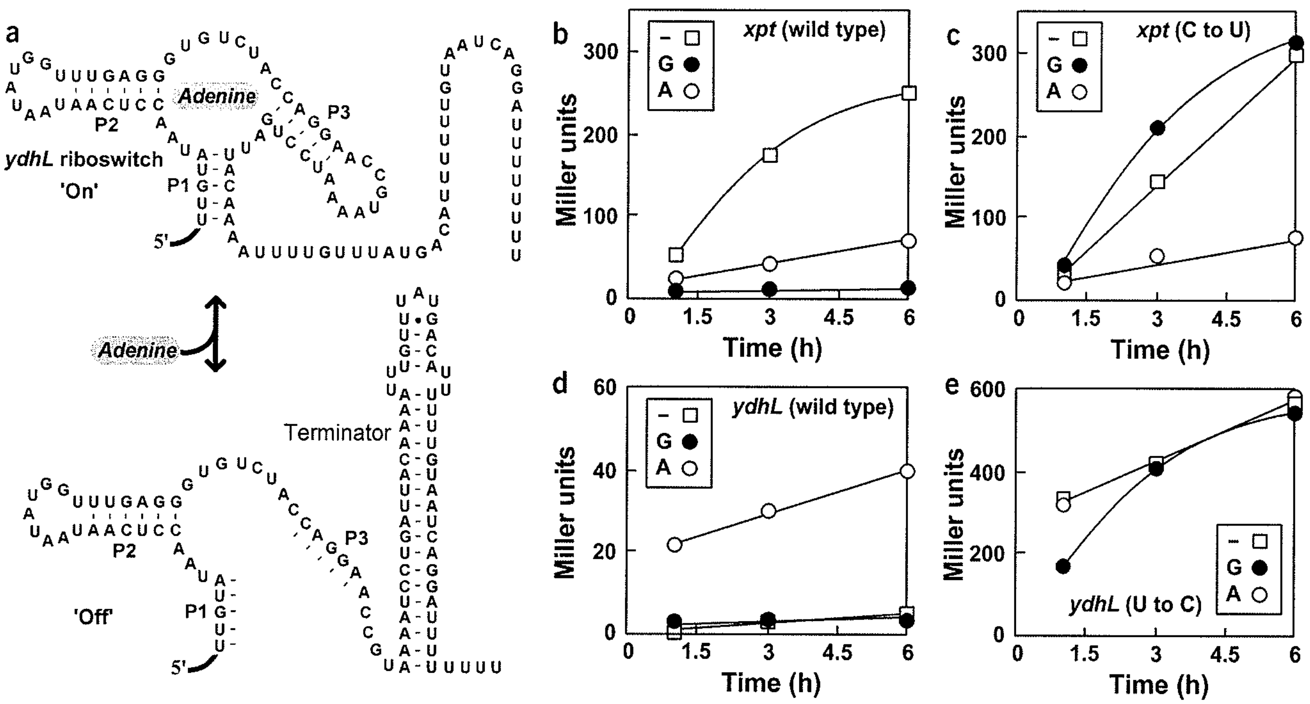


Figure 6 Model for the genetic control of *ydhL* by an adenine riboswitch and its function as a gene-activating element. (**a**) Sequence of the adenine riboswitch from *B. subtilis ydhL* and secondary structure models for the 'on' and 'off' states for gene regulation. (**b–e**) *In vivo* function of the wild-type *ydhL* riboswitch and of a variant form as determined by fusion to a β-galactosidase reporter gene.

mutations might fine-tune the binding affinity of the aptamers, whereas others could have resulted from neutral drift in the RNA sequence that is permitted because they retain the essential secondary-structure elements.

Genetic control and function of the *ydhL* mRNA

Mutant strains of *B. subtilis* that resist the toxic effects of 2-fluoroadenine were reported recently[9]. These mutations, which cause overexpression of the *ydhL* gene product, were mapped to the adenine riboswitch domain. In both cases, the changes (deletions) are expected to disrupt riboswitch function by eliminating a portion of the terminator stem or by eliminating both the terminator stem and portions of the adenine aptamer domain. In both cases, the variants prevent the riboswitch from adopting its default sate (transcription termination), causing unmodulated activation of gene expression.

The protein product of the *ydhL* gene (also called *pbuE*) has been proposed to be a purine efflux pump[9]. Thus the resistance to 2-fluoroadenine conferred on the cell by disruption of the adenine riboswitch from *ydhL* might be due to excretion of this toxic compound. We speculate that, in the wild-type organism, the presence of excess adenine within the cell probably induces increased expression of the *ydhL* gene to produce the purine efflux protein. Greater amounts of this protein then work to normalize the concentration of purines by pumping out of the cell one or more forms of this compound class.

Genetic modulation by rising metabolite concentrations

The adenine riboswitch from *B. subtilis* is also notable for its mechanism of action. In the majority of riboswitches examined so far, gene expression decreases as a result of metabolite binding. This occurs either by ligand-mediated formation of a terminator stem to prevent transcription of the complete mRNA, or by sequestration of the Shine-Dalgarno sequence and prevention of translation initiation[2]. In most instances, the downregulation of gene expression is expected, as a buildup of sufficient amounts of a particular metabolite would logically provide a signal to turn off genes involved in biosynthesis or import of the compound[18].

The adenine riboswitch from *ydhL* (and, presumably, the *add* riboswitches) belong to a group of genes whose functions require riboswitch activation in the presence of high concentrations of target compounds. For *ydhL*, disposal of excess purines would be important because certain purines such as guanine are insoluble at moderate concentrations. Alternatively, adenine deaminase would not need to be expressed if adenine concentrations were exceptionally low, and therefore we expect that the riboswitches from the *add* genes of *C. perfringens* and *V. vulnificus* might also be activated by ligand binding. Notably, T box domains, which are 5′ UTR structures that control the expression of many aminoacyl-tRNA synthetases in *B. subtilis* and other Gram-positive organisms[19], also induce gene expression in response to rising concentrations of the target they sense. However, unlike known metabolite-binding riboswitches, T box domains sense the biochemical precursor (nonaminoacylated tRNAs) to the products of the enzymes whose expression they control[20].

Although we expect riboswitches that induce gene activation in response to increasing metabolite concentration to occur less frequently owing to genetic necessity, there is no inherent reason why RNA folding would favor this distribution between gene-activating and gene-deactivating riboswitches. Regardless of whether the riboswitch responds to ligand binding by activating or repressing gene expression, the RNAs will exploit allosteric changes in secondary and/or tertiary structure that are based on the same principles of RNA folding. The only obligate difference between riboswitches that activate genes and those that repress them is in the fine structure of the expression platform. The aptamer domain can remain largely unchanged regardless of the genetic control mechanism; this accounts for their widespread sequence conservation in biological systems.

METHODS

Purine analogs. Guanine, adenine, 2,6-diaminopurine, 2-aminopurine, hypoxanthine, xanthine, 1-methyladenine, purine, 6-methylaminopurine, N^6-N^6 dimethyladenine, 6-mercaptopurine, 3-methyladenine, guanine-8[^{3}H] and adenine-2,8[^{3}H] were purchased from Sigma. 6-cyanopurine and 8-azaadenine were obtained from Aldrich and 2-chloroadenine, 8-chloroadenine from Biolog Life Science Institute.

DNA oligonucleotides. Oligonucleotides were synthesized by the Howard Hughes Medical Institute (HHMI) Keck Foundation Biotechnology Resource Center at Yale University (New Haven), purified by denaturing PAGE, and eluted from the gel by crush-soaking in a buffer containing 10 mM Tris-HCl (pH 7.5 at 23 °C), 200 mM NaCl and 1 mM EDTA. DNAs were precipitated with ethanol, resuspended in deionized water and stored at –20 °C until use.

In-line probing of RNA constructs. RNA constructs were synthesized from the corresponding PCR DNA templates by transcription *in vitro* using T7 RNA polymerase, dephosphorylated, and 5′-end labeled with ^{32}P as described[1]. In a typical in-line probing assay, 2 nM of labeled RNA was incubated for 40 h at 25 °C in a buffer containing 20 mM $MgCl_2$, 50 mM Tris-HCl (pH 8.3 at 25 °C) and 100 mM KCl in the absence or presence of purine compounds as indicated for each experiment. Purine concentrations ranging from 1 nM to 10 μM were used unless otherwise noted. At the end of each incubation, spontaneously cleaved products were separated using denaturing (8 M urea) 10% (w/v) PAGE, visualized with a PhosphorImager and quantified with ImageQuaNT (Molecular Dynamics).

Equilibrium dialysis. Equilibrium dialysis assays were conducted using a DispoEquilibrium Dialyzer (Harvard Biosciences) in which chambers A and B are separated by a 5,000 MWCO membrane. Chamber A contained 30 μl of [^{3}H]guanine (0.27 μCi) or [^{3}H]adenine (0.73 μCi) at a concentration of 100 nM in a buffer containing 50 mM Tris-HCl (pH 8.5 at 25 °C), 20 mM $MgCl_2$ and 100 mM KCl. A 30 μl aliquot of this buffer containing 3 μM RNA was delivered into chamber B. Equilibrations proceeded for 10 h at 25 °C. Subsequently, 5 μl was withdrawn from each chamber and quantified by liquid scintillation counting.

Construction of *xpt-* and *ydhL-lacZ* fusions. A DNA construct comprising nucleotides –468 to +9 relative to the translational start site of *ydhL* was amplified by PCR from *B. subtilis* strain 1A40 (Bacillus Genetic Stock Center, Columbus, Ohio, USA) with primers that introduced *Eco*RI-*Bam*HI restriction sites. The wild-type construct was inserted into pDG1661 at *Eco*RI-*Bam*HI restriction sites directly upstream of the *lacZ* reporter gene and sequenced to confirm its integrity. The resulting plasmid was used as a template for site-directed mutagenesis via the QuikChange kit (Stratagene) using the appropriate primer. Preparation of the wild-type *xpt-lacZ* construct was reported elsewhere[1]. Plasmid variants were integrated into the *amyE* locus of *B. subtilis* strain 1A40 and the transformants were confirmed as described[1].

***In vivo* analysis of riboswitch function.** Transformed *B. subtilis* cells were grown to mid log phase with constant shaking at 37 °C in minimal medium containing 0.4% (w/v) glucose, 20 g l^{-1} $(NH_4)_2SO_4$, 25 g l^{-1} K_2HPO_4, 6 g l^{-1} KH_2PO_4, 1 g l^{-1} sodium citrate, 0.2 g l^{-1} $MgSO_4 \cdot 7H_2O$, 0.2% (w/v) glutamate, 5 μg ml^{-1} chloramphenicol, 50 μg ml^{-1} L-tryptophan, 50 μg ml^{-1} L-lysine and 50 μg ml^{-1} L-methionine. Guanine or adenine was added to a final concentration of 0.1 mg ml^{-1}. Cells at mid exponential stage were harvested and resuspended in minimal medium in the presence or absence of purines and grown for an additional time as indicated for each experiment, at which time 1 ml of cell culture was assayed for β-galactosidase activity using a variation of the method described by Miller[21].

ACKNOWLEDGMENTS
We thank members of the Breaker laboratory for helpful discussions and especially W.C. Winkler for helpful comments on the manuscript. This work was supported by grants from the US National Institutes of Health (GM 559343; NHLBI-N01-HV-28186) and the US National Science Foundation (EIA-0129939; EIA-0323510; EIA-0324045). R.R.B is also grateful for support from the Hereditary Disease Foundation and from the David and Lucile Packard Foundation.

COMPETING INTERESTS STATEMENT
The authors declare that they have no competing financial interests.

Received 29 August; accepted 10 November 2003
Published online at http://www.nature.com/natstructmolbiol/

1. Mandal, M., Boese, B., Winkler, W.C. & Breaker, R.R. Metabolite-sensing riboswitches control fundamental biochemical pathways in bacteria. *Cell* **113**, 577–586 (2003).
2. Winkler, W.C. & Breaker, R.R. Genetic control by metabolite-binding riboswitches. *Chembiochem* 4, 1024–1032 (2003).
3. Sudarsan, N., Wickiser, J.K. Nakamura, S., Ebert, M.S. & Breaker, R.R. An mRNA structure that controls gene expression by binding lysine. *Genes Dev.* **17**, 2685–2697 (2003).
4. Winkler, W.C., Cohen-Chalamish, S. & Breaker, R.R. An mRNA structure that controls gene expression by binding FMN. *Proc. Natl. Acad. USA* **99**, 15908–15913 (2002).
5. Winkler, W.C., Nahvi, A., Sudarsan, N., Barrick, J.E. & Breaker, R.R. An mRNA structure that controls gene expression by binding *S*-adenosylmethionine. *Nat. Struct. Biol.* **10**, 701–707 (2003).
6. Nahvi, A. *et al.* Genetic control by a metabolite binding mRNA. *Chem. Biol.* **9**, 1043–1049 (2002).
7. Winkler, W., Nahvi, A. & Breaker, R.R. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature* **419**, 952–956 (2002).
8. Cristiansen, L.C., Schou, S., Nygaard, P. & Saxild, H.H. Xanthine metabolism in *Bacillus subtilis*: characterization of the *xpt-pbuX* operon and evidence for purine- and nitrogen-controlled expression of genes involved in xanthine salvage and catabolism. *J. Bacteriol.* **179**, 2540–1550 (1997).
9. Johansen, L.E., Nygaard, P., Lassen, C., Agersø, Y. & Saxild, H.H. Definition of a second *Bacillus subtilis pur* regulon comprising the *pur* and *xpt-pbuX* operons plus *pbuG*, *nupG* (*yxjA*), and *pbuE* (*ydhL*). *J. Bacteriol.* **185**, 5200–5209 (2003).
10. Soukup, G.A. & Breaker, R.R. Relationship between internucleotide linkage geometry and the stability of RNA. *RNA* **5**, 1308–1325 (1999).
11. Soukup, G.A., DeRose, E.C., Koizumi, M. & Breaker, R.R. Generating new ligand-binding RNAs by affinity maturation and disintegration of allosteric ribozymes. *RNA* **7**, 524–536 (2001).
12. Peracchi, A., Beigelman, L., Usman, N. & Herschlag, D. Rescue of abasic hammerhead ribozymes by exogenous addition of specific bases. *Proc. Natl. Acad. Sci. USA* **93**, 11522–11527 (1996).
13. Wilson, K.S. & von Hippel, P.H. Transcription termination at intrinsic terminators: the role of the RNA hairpin. *Proc. Natl. Acad. Sci. USA* **92**, 8793–8797 (1995).
14. Gusarov, I & Nudler, E. The mechanism of intrinsic transcription termination. *Mol. Cell* 4, 495–504 (1999).
15. Mironov, A.S. *et al.* Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *Cell* **111**, 747–756 (2002).
16. McDaniel, B.A.M., Grundy, F.J., Artsimovitch, I. & Henkin, T.M. Transcription termination control of the S box system: direct measurement of *S*-adenosylmethionine by the leader RNA. *Proc. Natl. Acad. Sci. USA* **100**, 3083–3088 (2003).
17. Epshtein, V., Mironov, A.S. & Nudler, E. The riboswitch-mediated control of sulfur metabolism in bacteria. *Proc. Natl. Acad. Sci. USA* **100**, 5052–5056 (2003).
18. Rosenfeld, N., Elowitz, M.B. & Alon, U. Negative autoregulation speeds the response times of transcription networks. *J. Mol. Biol.* **323**, 785–793 (2002).
19. Grundy, F.J. & Henkin, T.M. The T box and S box transcription termination control systems. *Frontiers Biosci.* **8**, d20–31 (2003).
20. Grundy, F.J., Winkler, W.C. & Henkin, T.M. tRNA-mediated transcription antitermination *in vitro*: codon-anticodon pairing independent of the ribosome. *Proc. Natl. Acad. Sci. USA* **99**, 11121–11126 (2002).
21. Miller, J.H. *A Short Course in Bacterial Genetics* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, USA, 1992).

Vol 441|29 June 2006|doi:10.1038/nature04740

nature

Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch

Alexander Serganov[1], Anna Polonskaia[1], Anh Tuân Phan[1], Ronald R. Breaker[2] & Dinshaw J. Patel[1]

Riboswitches are metabolite-sensing RNAs, typically located in the non-coding portions of messenger RNAs, that control the synthesis of metabolite-related proteins[1–3]. Here we describe a 2.05 Å crystal structure of a riboswitch domain from the *Escherichia coli thiM*[4] mRNA that responds to the coenzyme thiamine pyrophosphate (TPP). TPP is an active form of vitamin B_1, an essential participant in many protein-catalysed reactions[5]. Organisms from all three domains of life[6–9], including bacteria, plants and fungi, use TPP-sensing riboswitches to control genes responsible for importing or synthesizing thiamine and its phosphorylated derivatives, making this riboswitch class the most widely distributed member of the metabolite-sensing RNA regulatory system. The structure reveals a complex folded RNA in which one subdomain forms an intercalation pocket for the 4-amino-5-hydroxymethyl-2-methylpyrimidine moiety of TPP, whereas another subdomain forms a wider pocket that uses bivalent metal ions and water molecules to make bridging contacts to the pyrophosphate moiety of the ligand. The two pockets are positioned to function as a molecular measuring device that recognizes TPP in an extended conformation. The central thiazole moiety is not recognized by the RNA, which explains why the antimicrobial compound pyrithiamine pyrophosphate targets this riboswitch and downregulates the expression of thiamine metabolic genes. Both the natural ligand and its drug-like analogue stabilize secondary and tertiary structure elements that are harnessed by the riboswitch to modulate the synthesis of the proteins coded by the mRNA. In addition, this structure provides insight into how folded RNAs can form precision binding pockets that rival those formed by protein genetic factors.

More than 2% of the genes in some species are regulated by riboswitches[4,10–13], whose representative classes compete in number with known metabolite-sensing regulatory proteins[14]. On metabolite docking, the sensing domain of the riboswitch shows conformational stabilization, which typically alters the base-pairing arrangements in the adjoining expression platform carrying gene-expression signals[1–3]. Riboswitches have a remarkable affinity for their cognate ligands and can discriminate against even closely related analogues, as shown by biochemical experiments[1,13,15] and X-ray structures of purine-sensing riboswitches bound to hypoxanthine[16], guanine[17] and adenine[17]. TPP and pyrithiamine pyrophosphate (PTPP; Fig. 1a) are chemically more diverse than purines and pose several additional challenges for negatively charged RNA receptors, namely their large size, conformational flexibility and, most importantly, the presence of negatively charged phosphate groups[18].

We crystallized an 80-nucleotide TPP-sensing domain from the *thiM* mRNA[4]. This RNA conforms well with the consensus sequence and secondary structure model for TPP riboswitches that was established by previous phylogenetic[6–9] and biochemical[4,6,19] analyses (Supplementary Fig. S1). The structure (Fig. 1b–d, Supplementary

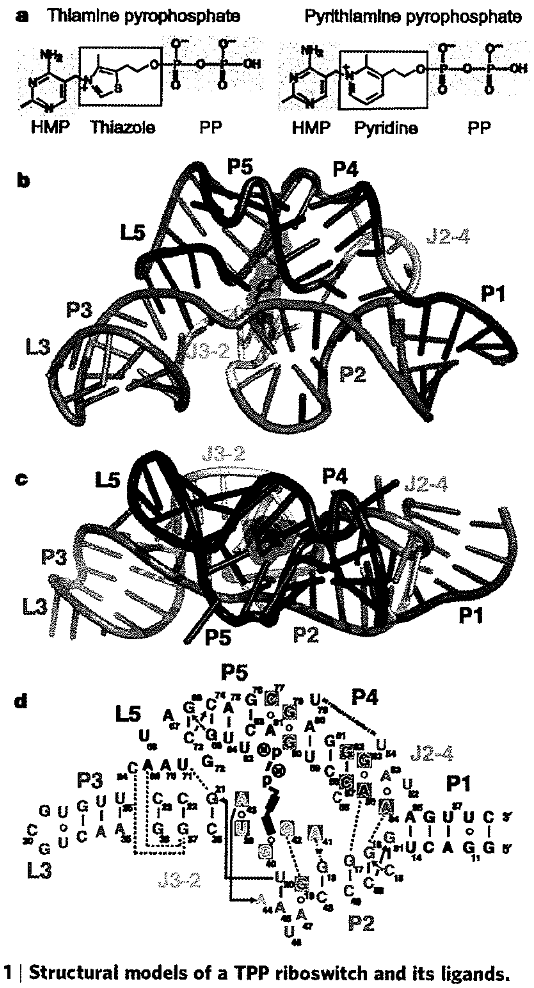


Figure 1 | Structural models of a TPP riboswitch and its ligands. **a**, Chemical structures of the natural metabolite TPP and the antimicrobial compound PTPP. **b**, **c**, Crystal structure of the TPP-bound sensing domain, showing front (**b**) and top (**c**) views. The RNA is in a stick-and-ribbon representation, with bound TPP in red. Stems, loops and junctions are colour coded. **d**, Schematic depiction of the RNA tertiary fold observed in the structure. Tertiary contacts formed by hydrogen bonds (w, water-mediated bonds) between bases and stacking interactions are represented by thin and thick dashed lines, respectively. Red shading shows nucleotides conserved in more than 97% of sequences (J. E. Barrick and R.R.B., unpublished observations). Encircled M notations represent Mg^{2+} ions.

[1]Structural Biology Program, Memorial Sloan-Kettering Cancer Center, New York, New York 10021, USA. [2]Department of Molecular, Cellular and Developmental Biology and Howard Hughes Medical Institute, Yale University, New Haven, Connecticut 06520, USA.

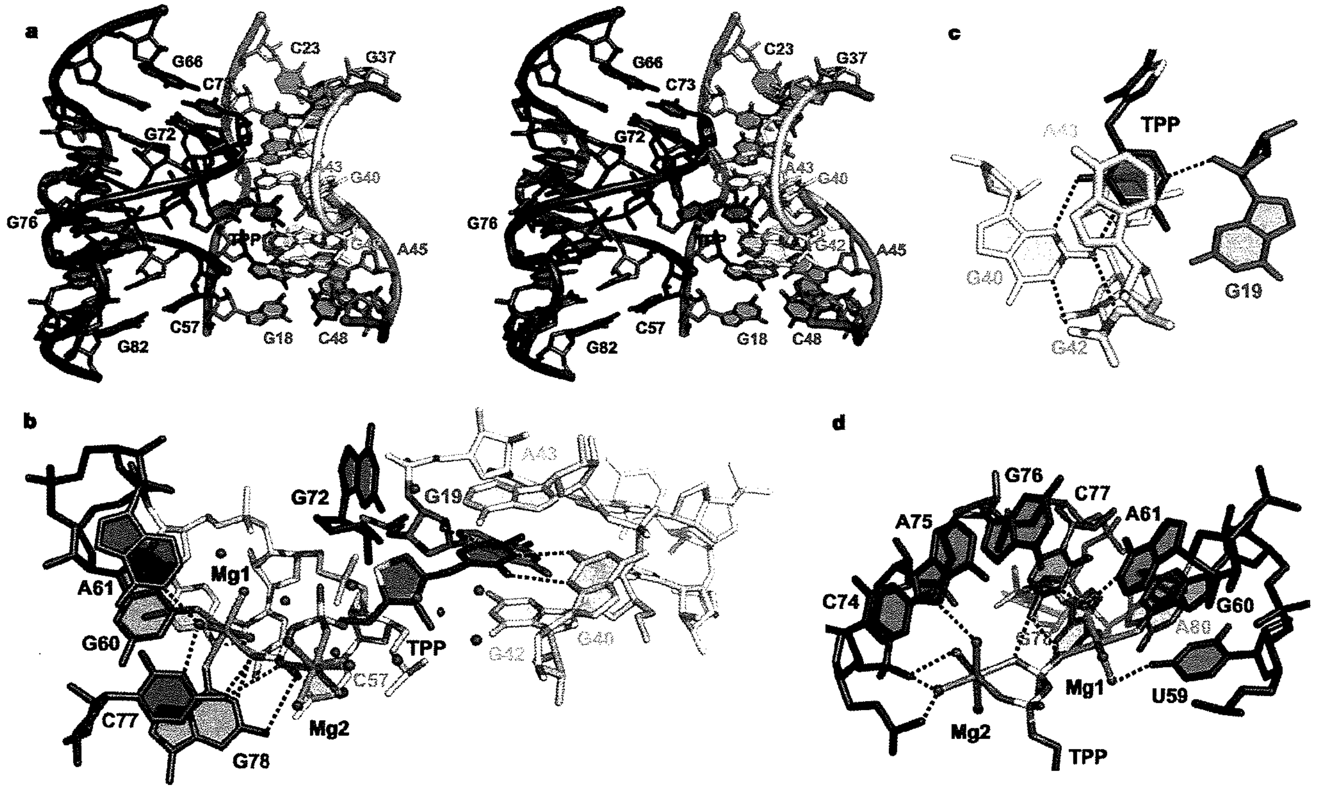


Figure 2 | Structure and interactions in the TPP-binding pocket. a, Stereo view of the central region of the complex containing bound TPP. **b,** View of TPP, coordinated Mg^{2+} ions (magenta) and water (blue spheres) in the binding pocket. **c,** Details of the interactions between the HMP ring and RNA. **d,** Hydrogen bonding between Mg^{2+} ions and RNA.

Fig. S2 and Supplementary Table S1) reveals that the conserved nucleotides and secondary structure elements common to all TPP riboswitches form a complex tertiary architecture consisting of two parallel helical domains (P2/J3-2/P3/L3 and J2-4/P4/P5/L5) connected to a helix (P1) by means of a three-way junction.

Unlike most other ligand–RNA interactions that exploit the helical grooves of RNA, TPP is positioned perpendicular to the two main helical domains and uses separate pockets formed by each helical segment to grip the ends of the compound in an extended conformation (Fig. 2). In addition to the interdomain bridge formed by the ligand, the riboswitch uses tertiary contacts between L5 and P3, and between J2-4, P2 and P4, to assist in stabilizing the global fold.

The J3-2 region (Fig. 1d) is essential in the recognition of the 4-amino-5-hydroxymethyl-2-methylpyrimidine (HMP) ring. The ring intercalates between G42 and A43, and forms hydrogen bonds between its polar functionalities, previously shown to be critical for molecular recognition[4], and the G40 base and the 2-OH′ of G19 (Fig. 2b, c). This binding arrangement is held by a T-loop-like turn[20] formed by the conserved U39-G40-A41-G42-A43 segment, closed by a reverse Hoogsteen U39•A43 pair (Fig. 3a). The backbone is extended at both G40-A41 and G42-A43, resulting in the mutual intercalation and insertion of the HMP ring between these interacting segments. Further, the fold is stabilized by continuous stacking of A41, G42, HMP, A43 and G21, and by the formation of (G19•A47)•G42 and (G18-C48)•A41 triples (Fig. 2a and Supplementary Fig. S3).

TPP riboswitches were the first of several classes found to make productive interactions with negatively charged phosphate groups[4,12,15]. The pyrophosphate group of TPP is bound in a spacious pocket by a pair of hexa-coordinated Mg^{2+} ions (Mg1 and Mg2) with octahedral ligation geometry that are positioned by direct and water-mediated hydrogen bonds with RNA (Fig. 2d and Supplementary Fig. S1). The terminal phosphate of TPP is coordinated to both Mg1 and Mg2, whereas the thiazole-linked phosphate is coordinated to Mg2. Mg1 is directly coordinated to the O^6 carbonyls of the conserved G60 and G78 that are located in the region previously implicated in pyrophosphate recognition[4] (Fig. 2b, d, and Supplementary Fig. S2). In addition, the conserved C77 and G78 form hydrogen bonds to the oxygen atoms of the terminal phosphate of TPP.

An analysis of 72 structures of proteins bound to TPP or its analogues from the Protein Data Bank indicates that proteins typically use charged amino acids to position Mg^{2+}, Ca^{2+} or Mn^{2+} ions in a site equivalent to Mg2. However, the Mg1 ion is unique to TPP riboswitches. A bivalent cation at this location allows TPP to reach into the pyrophosphate-binding pocket, thereby stabilizing tertiary interactions important for gene regulation and corroborating the requirement of Mg^{2+} for TPP binding in eukaryotic[21] and bacterial (Supplementary Fig. S4) TPP riboswitches. One could consider the ligand as TPP with two bound Mg^{2+} ions, which overcomes the ligand's strongly negative electrostatic character. Other riboswitches could use similar simple structural arrangements in combination with cations and water to bind phosphorylated ligands selectively, which would make RNA surprisingly adept at binding negatively charged ligands. The dual bivalent cation arrangement is also distinct from that found in ribosomal RNAs that bind non-phosphorylated antibiotics, in which some contacts with RNA are mediated by single cations[22–24].

Nucleotides distal to the ligand-binding sites, conserved in sequence or secondary structure, are important in forming the overall architecture of the TPP-sensing domain. J2-4 (Fig. 1d) is stabilized through the formation of a pair of stacked tetrads, where highly conserved A56•G83 and A53•A84 non-canonical pairs (Supplementary Fig. S5a, b) are aligned in the minor groove of adjacent G•C pairs within P2, forming type I A-minor motifs[25]. Continuous stacking is observed between P1 and P2, and between non-canonical

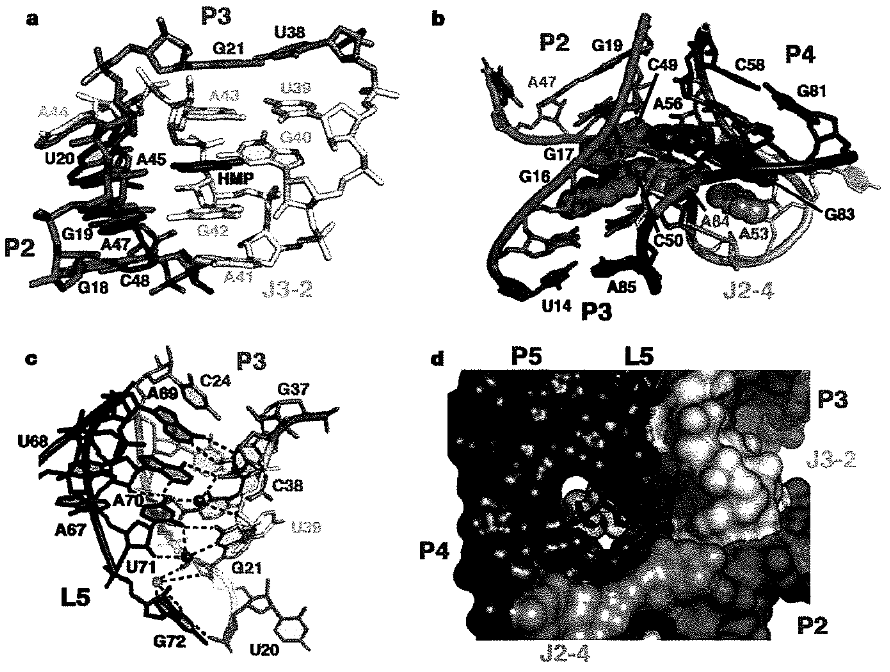


Figure 3 | Tertiary interactions defining TPP riboswitch structure and accessibility to the binding pocket. **a**, Interaction between J3/2 and P2, mediated by the HMP ring. **b**, Stabilization of the J2-4 junction by two stacked tetrads (in space-filling representation). **c**, Interactions between L5 and P3 mediated by three K^+ ions (red spheres). **d**, Surface representation of RNA and accessibility to the TPP-binding pocket. TPP is depicted in a stick and mesh representation.

pairs in J2-4 and P4 (Fig. 3b). This structural arrangement is expected to be stabilized by ligand binding, and reinforces previous suggestions[4,12,13] that stabilization of P1 stems in riboswitches is important for the control of gene expression.

Another key tertiary interface involves multiple base–base and base–backbone interactions between nucleotides of L5 and P3 (Fig. 3c), thereby locking into register side-by-side arrangements of P4/P5 and P2/J3-2/P3 segments (Fig. 1b). Three adjacent K^+ ions additionally stabilize the interactions (Fig. 3c). It should be noted that nucleotides of the L5/P3 interface are poorly conserved, and other TPP riboswitches could be stabilized by alternative tertiary contacts spanning this region.

A surface view of the complex shows that the thiazole and diphosphate moieties of TPP are visible, whereas the intercalated HMP ring is buried (Fig. 3d). The deep insertion of the HMP ring most probably implies a conformational transition within the riboswitch upon binding TPP. Support for this interpretation comes from in-line probing experiments, which have identified numerous internucleotide linkages that become conformationally restricted on TPP binding[4]. The structure shows that these regions either contact the ligand directly or participate in critical contacts that form the global structure of the sensing domain.

To assess global conformational changes brought about by ligand binding, we conducted nuclease V1 (helix-specific) and nuclease T2 (single-strand-specific) partial digests of the free and ligand-bound riboswitches. In the absence of TPP these nucleases cleave regions that participate in TPP binding, such as nucleotides (nt) 39–44 (J3-2) and nt 59–61 (part of P4/P5), as well as some regions that are not directly involved in formation of the TPP-binding pocket, such as nt 67–72 (L2) and nt 84–86 (Fig. 4 and Supplementary Figs S6 and S7). The addition of TPP decreases nuclease cleavage in regions that are important for both ligand binding and RNA folding (Fig. 4a, c). These results indicate that TPP not only stabilizes the two sections of the binding pocket by bridging the J3-2 and P4/P5 regions but also promotes the formation and/or stabilization of distal tertiary contacts within the J2-4/P2/P4 region and between L5 and P3.

Tertiary contacts associated with the TPP-stabilized fold of the riboswitch are also apparent in primer extension assays using reverse transcriptase (RT). In the presence of various analogues, including thiamine monophosphate (TMP), RT does not alter its pattern of pausing compared with that observed in the absence of ligand (Fig. 4b). This probably reflects the inability of TMP to span the two ligand-binding portions of the sensing domain fully and to stabilize the global fold, consistent with the progressive loss of binding affinity for thiamine ligands that carry one or no phosphates[4]. However, in the presence of TPP, RT pauses at two distinct positions (Fig. 4b). One pause at G86 occurs close to a pair of stacked tetrads anchoring tertiary contacts involving J2-4, P2 and P4, indicating that this conserved structural element becomes stabilized on ligand binding (Fig. 4c). A second RT pause occurs at C74 adjacent to the tertiary contact between L5 and P3, which seems to be sufficiently strong despite the disruption of base pairing in P1, P4 and P5 during primer extension.

The composite TPP-binding pocket can be defined as a molecular ruler in which only a ligand of proper length can stabilize tertiary interactions and modulate gene expression. This structural characteristic allows the riboswitch to achieve up to about 3,000-fold discrimination against shorter TMP and thiamine, and to provide feedback gene regulation only in the presence of biologically active TPP (ref. 4). In contrast to such unprecedented discrimination of related ligands by the RNA distance ruler, purine riboswitches have developed a completely different approach to distinguishing between related ligands. Adenine-specific and guanine-specific riboswitches use a single tight pocket to position their ligands precisely for Watson–Crick pairing with the discriminatory nucleotide, uracil or cytosine, respectively, thus making a single nucleotide the key element governing riboswitch regulation[16,17].

Two predominant modes of TPP-dependent gene regulation, translation inhibition and transcription termination, are illustrated in Fig. 4d. Both require an over-threshold concentration of TPP in the cell for interaction with the riboswitch[4,9,11,19]. As evident from our biochemical experiments, binding of TPP to the riboswitch induces

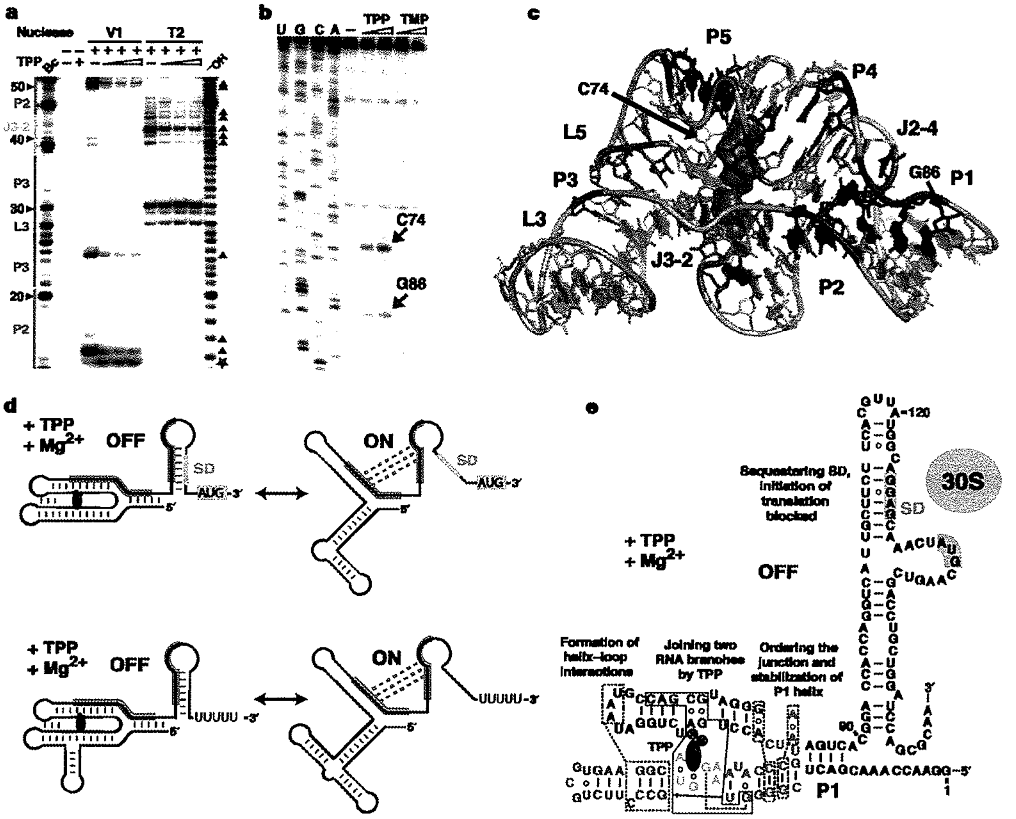


Figure 4 | Structural probing of the TPP riboswitch and implications for TPP-mediated gene repression. a, Representative RNase V1 and T2 cleavage patterns for the *thiM* riboswitch (nt 1–166). 5′ ^{32}P-labelled RNAs were treated by nucleases in the absence (−) or presence (+) of one, three and ten equivalents of TPP as described in Supplementary Methods and ref. 30. $^{-}$OH and Bc stand for ladders prepared by partial digestion with alkali or *B. cereus* RNase, respectively. The major cleavage protections and enhancements in the presence of TPP are labelled with triangles and stars, respectively. **b,** Primer extension analysis in the absence and presence of TPP or TMP. RT pauses are indicated by arrows. **c,** Summary of structure probing experiments. Major TPP protections against V1 (red) and T2 (magenta) RNases are shown. The RT pauses are indicated by arrows. **d,** Typical mechanisms of TPP-specific gene repression. Top: translation initiation regulation (*thiM* genes). Bottom: transcription termination regulation (*thiC* genes). Complementary sequences and alternate base-pairing are shown in blue. SD sequence and initiation codon are shaded green. TPP and Mg^{2+} ions are depicted in red and magenta, respectively. **e,** Diagram of the OFF state of the *E. coli thiM* riboswitch.

conformational rearrangements leading to stabilization of the overall RNA fold and, most importantly, to stabilization of conserved tertiary contacts adjacent to P1, as defined in the crystal structure (Fig. 4e). These interactions, in turn, stabilize P1 and promote folding of the expression platform, either to a hairpin that sequesters the Shine–Dalgarno (SD) sequence or to a terminator hairpin (OFF state); this results in the failure of translation initiation or premature transcription termination, respectively (Fig. 4d). Without TPP the riboswitch adopts alternative conformations, forming an anti-terminator hairpin or opening the SD sequence for ribosome binding (ON state).

Recent research[19] has revealed that the antimicrobial compound PTPP (Fig. 1a) binds to bacterial and fungal TPP riboswitches and can turn off the expression of critical biosynthetic genes. The structure indicates that the loss of PTPP activity, associated with drug-resisting mutations in TPP riboswitches, might be due to the disruption of key tertiary contacts made by tetrads (C50G, A84G, G86A) and J3-2 elements (A47Δ) (Fig. 3a, b). PTPP carries a pyridine ring in place of the thiazole ring of TPP (Fig. 1a). Because the TPP riboswitch does not make any substantive contacts with the ligand in this chemically distinctive region, our structure-based model reveals why PTPP functions as a mimic of the natural ligand (Supplementary Fig. S8). Given the important functional role of riboswitches in numerous microorganisms and the fact that riboswitches have not yet been detected in the human genome, structures of TPP and other riboswitch classes should enable researchers to employ rational drug discovery strategies to create novel classes of antibacterial and antifungal compounds that target riboswitches[19,26,27].

METHODS

Crystallization. The TPP–riboswitch complex was prepared by mixing RNA transcribed *in vitro* and TPP in a buffer containing 50 mM potassium acetate pH 6.9 and 5 mM $MgCl_2$. Crystals were grown by hanging-drop vapour diffusion. The complex solution and reservoir solution (28% w/v poly(ethylene glycol) 4000, 100 mM sodium acetate pH 4.8 and 200 mM ammonium acetate) were mixed in a 1:1 ratio and incubated at 20 °C. For soaking, crystals were washed with stabilizing solution lacking $MgCl_2$ and then incubated in the presence of 3 mM $Os(NH_3)_6$ for two days.

Structure determination. Native and $Os(NH_3)_6$ multiple anomalous diffraction (MAD) data were collected at beamline X25 at the Brookhaven National Synchrotron Light Source. Data were processed with the HKL2000 suite of programs (HKL Research). The structure was determined by using MAD osmium data and SOLVE/RESOLVE[28]. The RNA model was built using TURBO-FRODO (⟨http://afmb.cnrs-mrs.fr/rubrique113.html⟩) and refined with REFMAC[29] using a native data set (Supplementary Figs S9–S11). TPP and cations were added to the model on the basis of analysis of $2F_o - F_c$ and $F_o - F_c$ electron density maps. Na^+, K^+ and hydrated Mg^{2+} were modelled on the basis of the number of coordination bonds, their distances and their coordination geometry. RNA residue 55 has a partial electron density.

Received 16 February; accepted 22 March 2006.
Published online 21 May 2006.

1. Mandal, M. & Breaker, R. R. Gene regulation by riboswitches. *Nature Rev. Mol. Cell Biol.* 5, 451–463 (2004).

2. Nudler, E. & Mironov, A. S. The riboswitch control of bacterial metabolism. *Trends Biochem. Sci.* **29**, 11–17 (2004).
3. Soukup, G. A. & Soukup, J. K. Riboswitches exert genetic control through metabolite-induced conformational change. *Curr. Opin. Struct. Biol.* **14**, 344–349 (2004).
4. Winkler, W., Nahvi, A. & Breaker, R. R. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature* **419**, 952–956 (2002).
5. Schowen, R. L. in *Comprehensive Biological Catalysis* Vol. 2 (ed. Sinnott, M.) 217–266 (Academic, San Diego, 1998).
6. Sudarsan, N., Barrick, J. E. & Breaker, R. R. Metabolite-binding RNA domains are present in the genes of eukaryotes. *RNA* **9**, 644–647 (2003).
7. Kubodera, T. *et al.* Thiamine-regulated gene expression of *Aspergillius oryzae thiA* requires splicing of the intron containing a riboswitch-like domain in the 5′-UTR. *FEBS Lett.* **555**, 516–520 (2003).
8. Miranda-Rios, J., Navarro, M. & Soberón, M. A. A conserved RNA structure (*thi* box) is involved in regulation of thiamin biosynthetic gene expression in bacteria. *Proc. Natl Acad. Sci. USA* **98**, 9736–9741 (2001).
9. Rodionov, D. A., Vitreschak, A. G., Mironov, A. A. & Gelfand, M. S. Comparative genomics of thiamin biosynthesis in prokaryotes. New genes and regulatory mechanisms. *J. Biol. Chem.* **277**, 48949–48959 (2002).
10. Nahvi, A. *et al.* Genetic control by a metabolite binding mRNA. *Chem. Biol.* **9**, 1043–1049 (2002).
11. Mironov, A. S. *et al.* Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *Cell* **111**, 747–756 (2002).
12. Winkler, W. C., Cohen-Chalamish, S. & Breaker, R. R. An mRNA structure that controls gene expression by binding FMN. *Proc. Natl Acad. Sci. USA* **99**, 15908–15913 (2002).
13. Mandal, M., Boese, B., Barrick, J. E., Winkler, W. C. & Breaker, R. R. Riboswitches control fundamental biochemical pathways in *Bacillus subtilis* and other bacteria. *Cell* **113**, 577–586 (2003).
14. Winkler, W. C. Metabolic monitoring by bacterial mRNAs. *Arch. Microbiol.* **183**, 151–159 (2005).
15. Winkler, W. C., Nahvi, A., Roth, A., Collins, J. A. & Breaker, R. R. Control of gene expression by a natural metabolite-responsive ribozyme. *Nature* **428**, 281–286 (2004).
16. Batey, R. B., Gilbert, S. D. & Montagne, R. K. Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine. *Nature* **432**, 411–415 (2004).
17. Serganov, A. *et al.* Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs. *Chem. Biol.* **11**, 1729–1741 (2004).
18. Auffinger, P., Bielecki, L. & Westhof, E. Anion binding to nucleic acids. *Structure* **12**, 379–388 (2004).
19. Sudarsan, N., Cohen-Chalamish, S., Nakamura, S., Emilsson, G. M. & Breaker, R. R. Thiamine pyrophosphate riboswitches are targets for the antimicrobial compound pyrithiamine. *Chem. Biol.* **12**, 1325–1335 (2005).
20. Nagaswamy, U. & Fox, G. E. Frequent occurrence of the T-loop RNA folding motif in ribosomal RNAs. *RNA* **8**, 1112–1119 (2002).
21. Yamauchi, T. *et al.* Roles of Mg^{2+} in TPP-dependent riboswitch. *FEBS Lett.* **579**, 2583–2588 (2005).
22. Brodersen, D. E. *et al.* The structural basis for the action of the antibiotics tetracycline, pactamycin, and hygromycin B on the 30S ribosomal subunit. *Cell* **103**, 1143–1154 (2000).
23. Schlunzen, F. *et al.* Structural basis for the interaction of antibiotics with the peptidyl transferase centre in eubacteria. *Nature* **413**, 814–821 (2001).
24. Pioletti, M. *et al.* Crystal structures of complexes of the small ribosomal subunit with tetracycline, edeine and IF3. *EMBO J.* **20**, 1829–1839 (2001).
25. Nissen, P., Ippolito, J. A., Ban, N., Moore, P. B. & Steitz, T. A. RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. *Proc. Natl Acad. Sci. USA* **98**, 4899–4903 (2001).
26. Sudarsan, N., Wickiser, J. K., Nakamura, S., Ebert, M. S. & Breaker, R. R. An mRNA structure in bacteria that controls gene expression by binding lysine. *Genes Dev.* **17**, 2688–2697 (2003).
27. Hesselberth, J. R. & Ellington, A. D. A (ribo) switch in the paradigms of genetic regulation. *Nature Struct. Biol.* **9**, 891–893 (2002).
28. Terwilliger, T. C. & Berendzen, J. Automated MAD and MIR structure solution. *Acta Crystallogr. D* **55**, 849–861 (1999).
29. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53**, 240–245 (1997).
30. Serganov, A., Polonskaia, A., Ehresmann, B., Ehresmann, C. & Patel, D. J. Ribosomal protein S15 represses its own translation via adaptation of an rRNA-like fold within its mRNA. *EMBO J.* **22**, 1898–1908 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank L. Malinina and Y.-R. Yuan for extensive discussions; A. Kazantsev and M. Yusupov for the gift of $Os(NH_3)_6$; and M. Becker for assistance with synchrotron data collection on beamline X25 at the Brookhaven National Laboratory, supported by the US Department of Energy. D.J.P. and R.R.B. were supported by funds from the NIH.

Author Contributions A.S. was responsible for all structural and biochemical aspects of this project. A.P. participated in RNA preparation and crystallization trials, and A.T.P. participated in NMR experiments. D.J.P., A.S. and R.R.B. co-wrote the paper. All authors discussed the results and commented on the manuscript.

Author Information Coordinates of the X-ray structure of the TPP-bound riboswitch have been deposited in the RCSB Protein Data Bank under accession code 2GDI. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.S. (serganoa@mskcc.org) or D.J.P. (pateld@mskcc.org).

Chemistry & Biology, Vol. 11, 1729–1741, December, 2004, ©2004 Elsevier Ltd. All rights reserved. DOI 10.1016/j.chembiol.2004.11.018

Structural Basis for Discriminative Regulation of Gene Expression by Adenine- and Guanine-Sensing mRNAs

Alexander Serganov,[1] Yu-Ren Yuan,[1]
Olga Pikovskaya,[1] Anna Polonskaia,[1]
Lucy Malinina,[1] Anh Tuân Phan,[1]
Claudia Hobartner,[2] Ronald Micura,[2]
Ronald R. Breaker,[3] and Dinshaw J. Patel[1,*]
[1]Structural Biology Program
Memorial Sloan-Kettering Cancer Center
New York, New York 10021
[2]Institute for Organic Chemistry
Center for Molecular Biosciences
Leopold Franzens University
6020 Innsbruck
Austria
[3]Department of Molecular, Cellular,
and Developmental Biology
Yale University
New Haven, Connecticut 06520

Summary

Metabolite-sensing mRNAs, or "riboswitches," specifically interact with small ligands and direct expression of the genes involved in their metabolism. Riboswitches contain sensing "aptamer" modules, capable of ligand-induced structural changes, and downstream regions, harboring expression-controlling elements. We report the crystal structures of the *add* A-riboswitch and *xpt* G-riboswitch aptamer modules that distinguish between bound adenine and guanine with exquisite specificity and modulate expression of two different sets of genes. The riboswitches form tuning fork-like architectures, in which the prongs are held in parallel through hairpin loop interactions, and the internal bubble zippers up to form the purine binding pocket. The bound purines are held by hydrogen bonding interactions involving conserved nucleotides along their entire periphery. Recognition specificity is associated with Watson-Crick pairing of the encapsulated adenine and guanine ligands with uridine and cytosine, respectively.

Introduction

RNA-mediated regulation [1–3] and catalysis [4, 5] are critically dependent on both nucleic acid architecture [6] and recognition [7, 8]. RNA aptamer-based systems serve as exceptional modules for ligand recognition and catalysis and exhibit tunable specificities and enantiomeric selectivities [9, 10]. Aptamer and catalytic RNA modular domains can be coupled through RNA linker elements to generate allosteric ribozymes [11]. The functional principle of such designed molecular switches reflects propagation of ligand-associated adaptive transitions in the aptamer domain to the ribozyme domain, which in turn can influence ribozyme-folding patterns. Such in vitro engineered tripartite RNA switches [12] have been successfully constructed to generate designed molecular sensors [13] and new families of genetic control elements [14].

The natural counterparts of in vitro engineered allosteric switches are recently discovered metabolite-sensing mRNAs that could potentially modulate the expression of many genes. A structure-function perspective of this new and unexpected role for mRNA could provide critical information for defining the molecular basis of allosteric mRNA transitions associated with the modulation of gene expression levels and metabolic homeostasis. Indeed, principles related to recognition of the aptamer scaffolds of metabolite-sensing mRNAs could underlie new approaches to drug design and to development of molecular sensors [15].

Recent reports describe the identification of RNA genetic control elements that bind coenzyme B_{12} [16], flavin mononucleotide [17, 18], thiamine pyrophosphate [17, 19], S-adenosylmethionine [20–22], guanine [23], adenine [24], lysine [25, 26], and glycine [27]. Interaction of these small organic molecules with mRNA domains, or riboswitches, modulates expression of the genes involved in metabolism of these compounds, accounting for approximately 2% of the bacterial genes [23]. The field of metabolite-sensing mRNAs continues to flourish as outlined in recent reviews [28, 29], with a recent entry being a glucosamine-6-phosphate-sensing mRNA, which introduces a new paradigm, namely that the riboswitch is also a ribozyme [30]. More generally, it is conceivable that yet-to-be-identified riboswitches could also control other RNA-associated processes, such as processing, transport, and degradation [28, 31]. Genetic control by metabolite-sensing mRNAs has also been recently extended to eukaryotes [32].

Bacterial riboswitches are typically positioned within the 5′-untranslated region of the mRNA under control. Riboswitches are composed of a ligand binding aptamer domain and an expression platform that interfaces with RNA elements involved in gene expression. The natural aptamers, which are highly conserved within organisms and among domains of life, form independently folded modular units, whereas the expression platforms vary in sequence, topology, and mechanism. The riboswitches can adopt two distinct conformations: the metabolite bound and metabolite-free folds, involving alternative base-pairing of the regulatory RNA region. If the specific ligand is present above the threshold concentration in the cell, it interacts with the aptamer domain and stabilizes the metabolite bound fold of the riboswitch, thereby preventing formation of the alternative conformation. This results in most cases in either stabilization or disruption of the regulatory hairpin within the expression platform, thereby influencing gene expression. The 70–170-nucleotide natural aptamer domains [29] are larger than their in vitro selected counterparts [9, 10], and the increased complexity and information content of the former may reflect their need to function as high-affinity and -selectivity RNA receptors. Indeed, the natural aptamers exhibit

*Correspondence: pateld@mskcc.org

binding affinities in the nM to low μM range, with the affinities decreasing from 10- to 100-fold on proceeding from the aptamer domain alone to the intact riboswitch. These numbers reflect the need for riboswitches to dynamically and rapidly regulate gene expression in both a temporal and a spatial manner, while retaining the selectivity of the response to a specific subset of the organism's genes.

A search was undertaken to identify purine-sensing mRNAs by examining the regulatory mechanisms of purine metabolism. These efforts focused on the *xpt-pbuX Bacillus subtilis* operon (hereafter termed *xpt*) that encodes genes involved in guanine metabolism and is regulated by certain purines, such as guanine, hypoxanthine, and xanthine [23]. A database search for evolutionary conservation within the 5′-UTR of *xpt* identified five transcriptional units with closely corresponding sequences and predicted secondary folds, termed the G-box. This domain is composed of three stems (P1, P2, and P3) connected by the junction, with significant conservation within the junction, hairpin loops (L2 and L3), and junctional P1 residues (Figure 1A; shown in red; left panel). In addition, the optimal lengths of stems P2 and P3 are 7 and 6 bp, respectively, and the matching sequences of the hairpin loops of L2 and L3 could allow potential pseudoknot formation [23]. An "in-line" probing assay, based on the patterns of spontaneous RNA cleavage, was used to establish that a 93-mer G-box *xpt* mRNA construct underwent a pronounced conformational transition on addition of guanine [23]. Guanine bound this RNA with approximately 5 nM affinity and reduced spontaneous cleavage throughout the junctional segment. Hypoxanthine and xanthine bound the xpt mRNA with 10-fold reduced affinity, whereas adenine binding was reduced by six orders of magnitude. Further, substantial loss of binding affinity was associated after alteration of every functional group on the guanine ring, suggesting that the guanine is completely encapsulated within the RNA fold in the complex [23].

It has been recently demonstrated that the *ydhL* gene, encoding for the putative purine efflux pump of *B. subtilis*, and the *add* gene, encoding for adenine deaminase from *Clostridium perfringens* and *Vibrio vulnificus*, harbor mRNA elements that sense adenine [24]. The secondary structure of the aptamer domain of the *ydhL* adenine-sensing mRNA (Figure 1B) is very similar to its *xpt* guanine-sensing mRNA counterpart. The *ydhL* mRNA binds most tightly to 2,6-diaminopurine (10 nM affinity) and less tightly (300 nM affinity) to adenine and 2-aminopurine. It also discriminates against guanine (>10,000 nM affinity) and purine (30,000 nM affinity). The guanine-sensing *xpt* and adenine-sensing *ydhL* mRNAs differ within the junction-connecting regions of their aptamer folds at three positions [24]. Two of these occur at positions that are known to be variable and are unlikely to influence ligand recognition. The third difference occurs at position 74, which is C in the guanine-sensing *xpt* mRNA and U in the adenine-sensing *ydhL* mRNA. Replacing this C with U in the *xpt* RNA alters its specificity from guanine to adenine; replacing the corresponding U with a C in the *ydhL* RNA alters its specificity from adenine to guanine [24]. Thus, remarkably, single-nucleotide substitutions can be used to interchange the guanine/adenine specificities within the aptamer domains.

Riboswitches appear to control gene expression by metabolite-modulated allosteric interconversions between alternate base-paired structures. The adaptive conformational transitions associated with metabolite binding to the aptamer domains are then harnessed through the expression platform. The *xpt* G-riboswitch has been shown to control gene expression through transcriptional termination [23]. Thus, the mRNA forms an antiterminator in the absence of guanine, thereby allowing RNA transcription elongation to proceed to completion (Figure 1A, right panel). The presence of guanine results in stabilization of the aptamer domain, thereby facilitating terminator formation and shutting down transcription (Figure 1A, left panel). In contrast, the *ydhL* A-riboswitch has been shown to control gene expression through transcriptional activation [24]. This mRNA forms a terminator in the absence of adenine (Figure 1B, right panel) but an antiterminator in the presence of adenine (Figure 1B, left panel). The *add* A-riboswitch does not have a stretch of uridines characteristic of transcriptional terminators but rather contains nonpaired Shine-Dalgarno GAA and initiation codon sequences, immediately downstream of the aptamer domain. Such sequences are likely to control gene expression through translational activation, whereby the Shine-Dalgarno and initiation codon sequences (both shaded in orange) are sequestered through pairing interactions in the absence of adenine (Figure 1C, right panel). The presence of adenine results in stabilization of the aptamer fold, thereby releasing these segments (Figure 1C, left panel) for interaction with ribosomal RNA and tRNA, resulting in initiation of translation. These results establish that the same mRNA aptamer fold can facilitate transcriptional termination and activation on the one hand and translational activation on the other, depending on the composition of the expression platform.

We now report the crystal structures of adenine bound to the aptamer domain of the adenine-sensing *add* mRNA and guanine bound to the aptamer domain of the guanine-sensing *xpt* mRNA. Our successful structure determination highlights the molecular principles by which a single nucleotide is capable of altering binding specificity [23, 24] and thereby switching gene expression patterns. Our structures complement the results in a just-published paper by Robert Batey and coworkers on the crystal structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine [33].

Results

Adenine- and Guanine-Riboswitch Complexes

Our structural studies were initially undertaken on the adenine-sensing *B. subtilis ydhL* and *V. vulnificus add* mRNA aptamer domains and the guanine-sensing *B. subtilis xpt* mRNA aptamer domain. The well-established *B. subtilis xpt* G-riboswitch and the *ydhL* A-riboswitch differ in over 20 nucleotide positions spanning the aptamer domain and are significantly different in the expression platform sequences, leading to either repression or activation of transcription, respectively (Figures 1A and

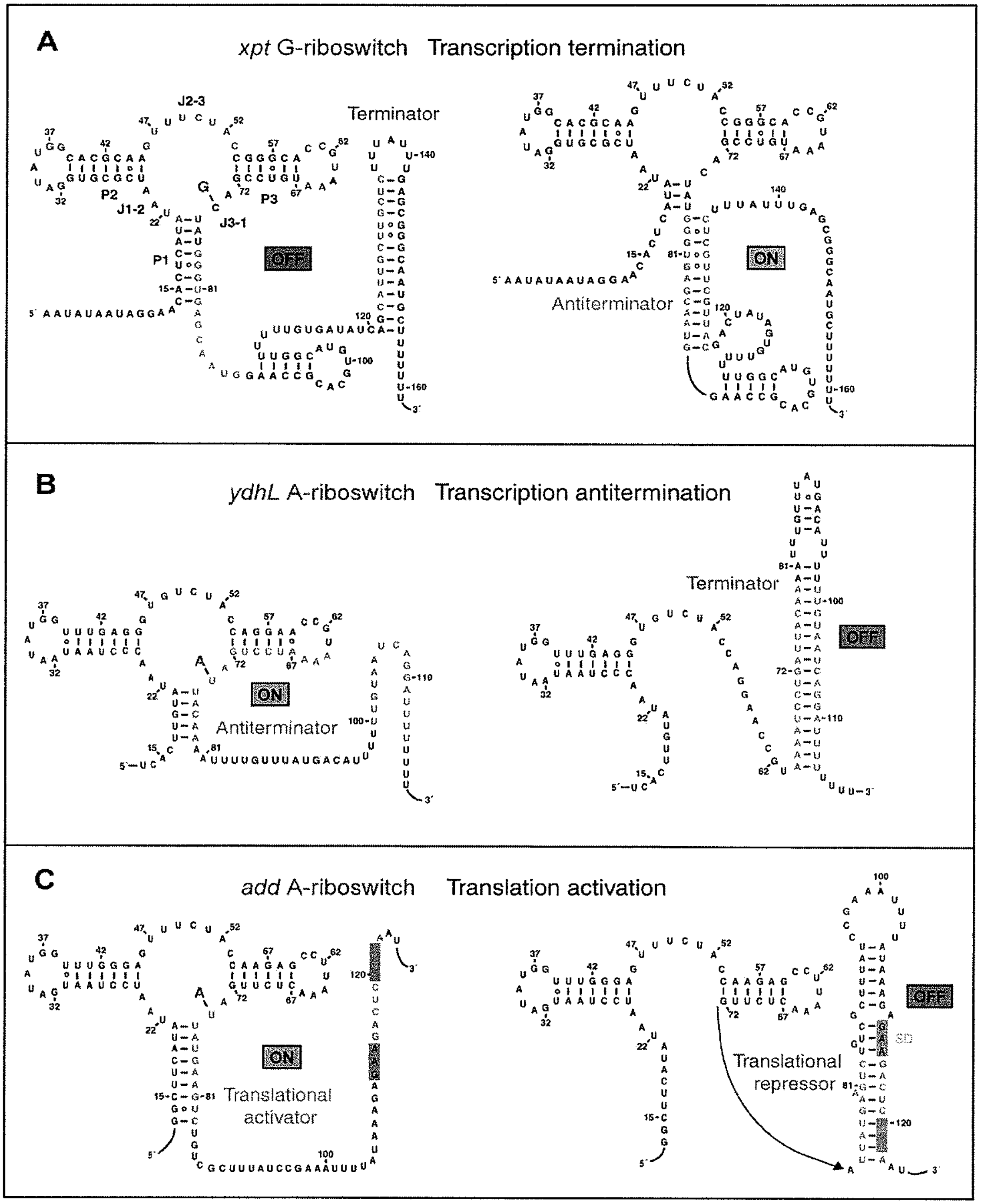


Figure 1. Mechanisms for Regulation of Gene Expression by Purine Riboswitches

(A) Schematic showing control of gene expression by the *xpt* G-riboswitch through transcriptional termination [23]. Highly conserved residues in the G-box of the aptamer domain are marked in red.

(B) Gene expression regulation by the *ydhL* A-riboswitch through disruption of the terminator hairpin [24].

(C) Control of gene expression by the *add* A-riboswitch most likely through translational activation. The Shine-Dalgarno GAA sequence and the initiation codon are both shaded in orange.

1B). Even though there are only a few differences between the adenine-sensing aptamer domains of the *V. vulnificus add* and *B. subtilis ydhL* A-riboswitches, their expression platforms are quite distinct, such that the former likely activates translation rather than transcription (Figure 1C).

We were only successful in growing diffraction quality crystals of adenine bound to the 71-mer adenine-

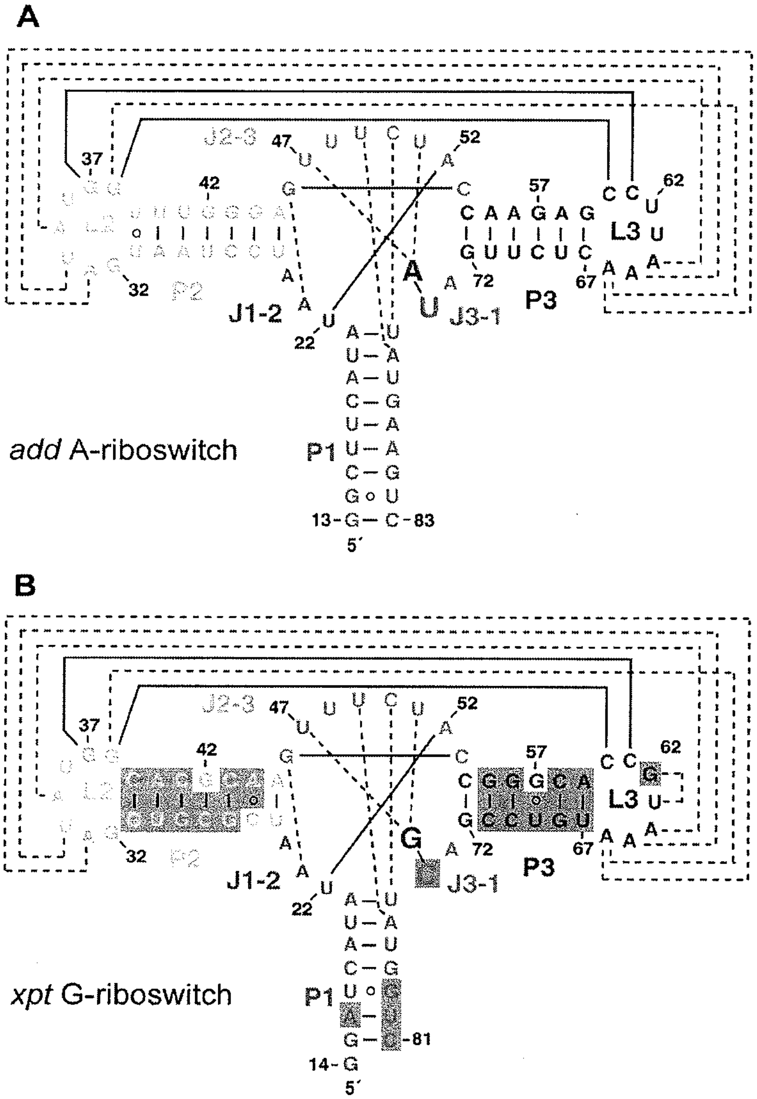


Figure 2. Sequence and Stem Secondary Structures of Aptamer Domains of Purine-Sensing mRNAs

(A) *V. vulnificus* 71-mer *add* A-riboswitch; (B) *B. subtilis* 68-mer *xpt* G-riboswitch. The color-coding scheme is as follows: Stems P1, P2, and P3 are green, yellow, and blue, respectively. Loops L2 and L3 are yellow and blue, respectively. Junction-connecting segments J1-2, J2-3, and J3-1 are cyan, orange, and violet, respectively. The specificity-determining pyrimidine residue at position 74 is highlighted with a larger lettering size. The bound adenine and guanine ligands are indicated in red lettering. Tertiary pairing alignments involving Watson-Crick and noncanonical base pairs are shown by full and dashed lines, respectively. The shaded regions in (B) highlight differences between the A- and G-riboswitches.

sensing *add* mRNA aptamer domain (hereafter designated A-riboswitch; Figure 2A) and guanine bound to the 68-mer guanine-sensing *xpt* mRNA aptamer domain (designated G-riboswitch; Figure 2B). We have adopted a simple nomenclature and color identification scheme for the aptamer domains, as outlined in the legend to Figure 2. Because the 5′ end of the *add* A-riboswitch was not experimentally determined, we have kept the *xpt* G-riboswitch numbering for both riboswitches.

Complex Formation Monitored by NMR

We have prepared the riboswitch RNAs by in vitro transcription and have incorporated a hammerhead ribozyme at the 3′ end to facilitate preparation of homogeneous transcripts. Binding of the ligand adenine and guanine to A- and G-riboswitches, respectively, has been monitored by nuclear magnetic resonance (NMR) spectroscopy. The imino proton NMR spectra (10.5–14.5 ppm) of the riboswitches in the absence and presence of one equivalent of bound ligand in 50 mM potassium acetate buffer (pH 6.8) are plotted in Figure 3. We observe exceptionally well-resolved imino proton (guanine N^1H and uridine N^3H) NMR spectra for both complexes (Figures 3B and 3D), consistent with formation of a single folded species in solution. Note the presence of new imino proton resonances between 13 and 14 ppm and between 10 and 11 ppm in these spectra. Low physiological concentration of Mg (2 mM) was required to drive complex formation to completion. In addition, exchange between free and bound states is slow on the NMR time scale, as reflected in doubling of resonances from free and bound imino protons that were observed on addition of substoichiometric (0.5 equivalents) of added adenine or guanine, characteristic of tight complex formation.

We have also investigated the binding of adenine and its analogs to the aptamer domain of the adenine-sensing *ydhL* mRNA. The imino proton spectra establish

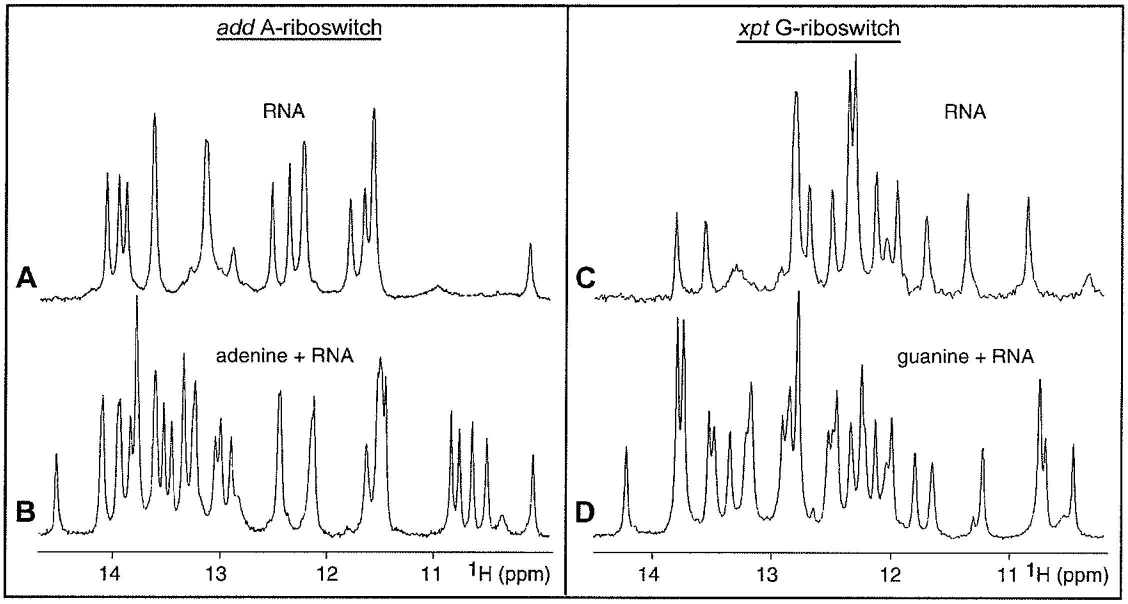


Figure 3. Imino Proton NMR Spectra of Purine Binding to Aptamer Domains of Purine Riboswitches

Imino proton NMR spectra (10–15 ppm) of the 71-mer *add* A-riboswitch in the absence (A) and presence (B) of one equivalent of adenine and of the 69-mer (with GGC and GUC sequences on the 5′ and 3′ ends, respectively) *xpt* G-riboswitch in the absence (C) and presence (D) of one equivalent of guanine. The spectra were recorded in 50 mM potassium acetate (pH 6.8) at 25°C. For both riboswitches, 2 mM Mg was added to drive complex formation to completion.

that this A-riboswitch forms 1:1 tight complexes not only with adenine, but also with 2-aminopurine and 2,6-diaminopurine (see Figure S1 in the Supplemental Data available with this article online). Complex formation is in slow exchange, with similar additional imino proton markers, characteristic of the bound state, observed in all three complexes.

Crystallization and Structure Determination

Despite sequence similarities, crystals of the A-riboswitch-adenine and G-riboswitch-guanine complexes were obtained under very different conditions. Crystals of the A-riboswitch complex were grown under high (200 mM) Mg concentration, pH 9.0 buffer, whereas the crystals of the G-riboswitch complex were grown under lower (20 mM) Mg concentration, pH 5.2 buffer. The A-riboswitch complex crystals belonged to space group $P2_12_12$ and diffracted to 2.1 Å, whereas the crystals of the G-riboswitch complex belonged to space group $C222_1$ and diffracted to 2.4 Å resolution.

We first solved the structure of the A-riboswitch-adenine complex. The native crystals of this complex were soaked in 60 mM $BaCl_2$ solution, and the structure was determined by using the anomalous properties of Ba to solve the phase problem (see Experimental Procedures for details). The structure contains all RNA residues, 5 Mg cations, and 66 water molecules in the asymmetric unit and was refined to final R-factor/R-free values of 23.1/29.7 (Table S1). The G-riboswitch-guanine complex structure was solved by molecular replacement with the structure of the A-riboswitch-adenine complex as a model. In addition to RNA residues, the structure contains 18 water molecules per RNA molecule and was refined to final R-factor/R-free values of 23.2/26.4 (Table S1).

add A-Riboswitch-Adenine Complex Structure

Overall Topology

The adenine bound A-riboswitch complex adopts a tuning fork-like compact fold (schematic in Figure 4A; structures in Figures 4B and 4C), where stem P1 forms the handle of the tuning fork, and stems P2 and P3, which form the prongs, are aligned parallel to each other and anchored at the tips through extensive interaction between their hairpin loops L2 and L3. The central internal bubble zippers up through stacked base triple alignments between the three junction-connecting segments, J1-2, J2-3, and J3-1, and two junctional base pairs of stem P1, thereby generating an adenine-sensing pocket within the resulting core segment of the RNA scaffold.

Zippering Up of Internal Bubble

The junctional bubble of the A-riboswitch contains trinucleotide J1-2, octanucleotide J2-3, and dinucleotide J3-1 junction-connecting segments (Figure 2A). Alignment of residues from all three junction-connecting segments on complex formation with adenine results in formation of a central core scaffold centered about the adenine binding site (stereo view in Figure 5A). Two triples, A23•(G46-C53) and water-mediated A73•(A52-U22), involving residues from J1-2, J2-3, and J3-1, are located above the adenine binding site, with adenines A23 and A73 positioned in the minor groove of their respective Watson-Crick base pairs (Figure 5B). Two additional triples, C50•(U75-A21) and U49•(A76-U20),

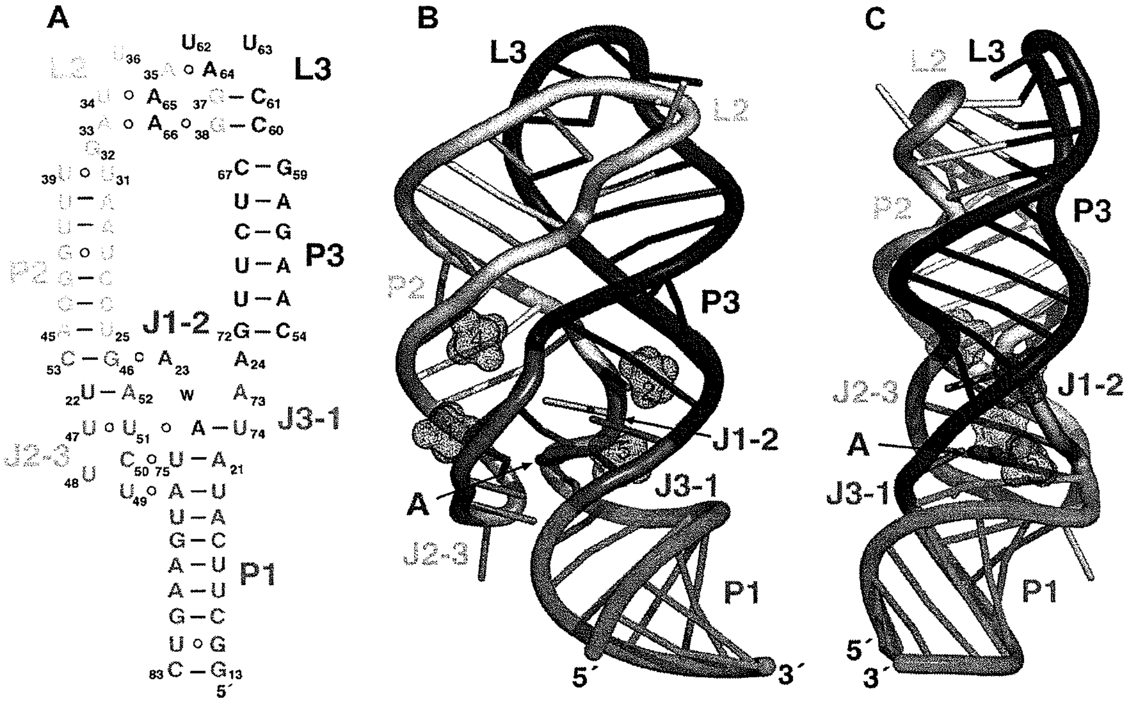


Figure 4. Schematic and Structure of the A-Riboswitch-Adenine Complex

(A) Schematic highlighting tertiary interactions in the folded structure of the A-riboswitch-adenine complex. The color-coding is as outlined in the caption to Figure 2.

(B and C) Ribbon representations (rotated by 90° along the vertical axis) of the A-riboswitch-adenine complex with the same color-coding scheme. The bound adenine is shown in red in a stick representation. Four of the five hydrated Mg cations are shown as dotted surfaces (remaining Mg is involved in packing interactions).

involving residues in J2-3 and junctional base pairs of stem P1, are located below the adenine binding site, with C50 and U49 also positioned in the minor groove of their respective Watson-Crick base pairs. Finally, adenine recognition occurs through formation of a U51•(adenine-U74) triple, involving residues in J2-3 and J3-1, with U51 positioned in the minor groove of its Watson-Crick base pair. The stacking patterns of the adenine-containing U51•(adenine-U74) triple with the flanking A73•(A52-U22) and C50•(U75-A21) triples are shown in Figures 5C and 5D, respectively. Thus, the compact core is composed of a five-tiered triplex with extensive base stacking between tiers. Among the remaining J2-3 residues, U47 is positioned in the groove and pairs with the bound adenine and U51, whereas U48 is directed outwards from the stacked base triple architecture.

Kissing Interaction between Hairpin Loops

Hairpin loops L2 and L3, each of which contains seven nucleotides, are anchored together through formation of 5 bp in the structure of the A-riboswitch-adenine complex (Figures 6A and 6B). Two of these pairs, G38-C60 and G37-C61, are aligned through Watson-Crick pairing (Figure 6C), and three others form noncanonical pairs [34, 35]. The latter include the *trans* U34•A65, *trans* A33•A66 pair and *trans* A35•A64 pairs, with pairing alignments shown in Figure 6C. In addition, further alignment amongst the pairs results in formation of the A33•A66•C60-G38 tetrad (Figure 6E) and the U34•A65•C61-G37 tetrad (Figure 6D) platforms, thereby additionally anchoring the kissing interaction between the hairpin loops. Among the remaining loop bases, G32 is stacked between A33 and the U31•U39 junctional pair of P2, U36 is directed outwards from the stacked array of kissing loop interactions, and U62 and U63 form the tip of the kissing loop scaffold, with U62 stacked over the A35•A64 pair (Figure 6C). The extensive kissing interaction between hairpin loops results in a parallel alignment of stems P2 and P3.

Adenine Recognition Specificity

The bound adenine moiety in the *add* A-riboswitch complex is held in position through formation of direct hydrogen bonds with three base edges and a sugar 2′-OH group (Figure 7A). The N^1 and N^6H_2 atoms positioned along the Watson-Crick edge of the bound adenine form a pair of hydrogen bonds with the Watson-Crick edge of U74 of J3-1. The N^3 and N^9H atoms positioned along the minor groove edge of the bound adenine form a pair of hydrogen bonds with the Watson-Crick base edge of U51 of J2-3, with N^9H also forming a hydrogen bond to O^2 of U47 of J2-3. The N^7 atom positioned along the major groove edge of the bound adenine forms a hydrogen bond with the 2′-OH of U22 of J1-2. Thus, all heteroatoms along the periphery of the bound adenine ring are recognized through hydrogen bond formation with residues from junction-

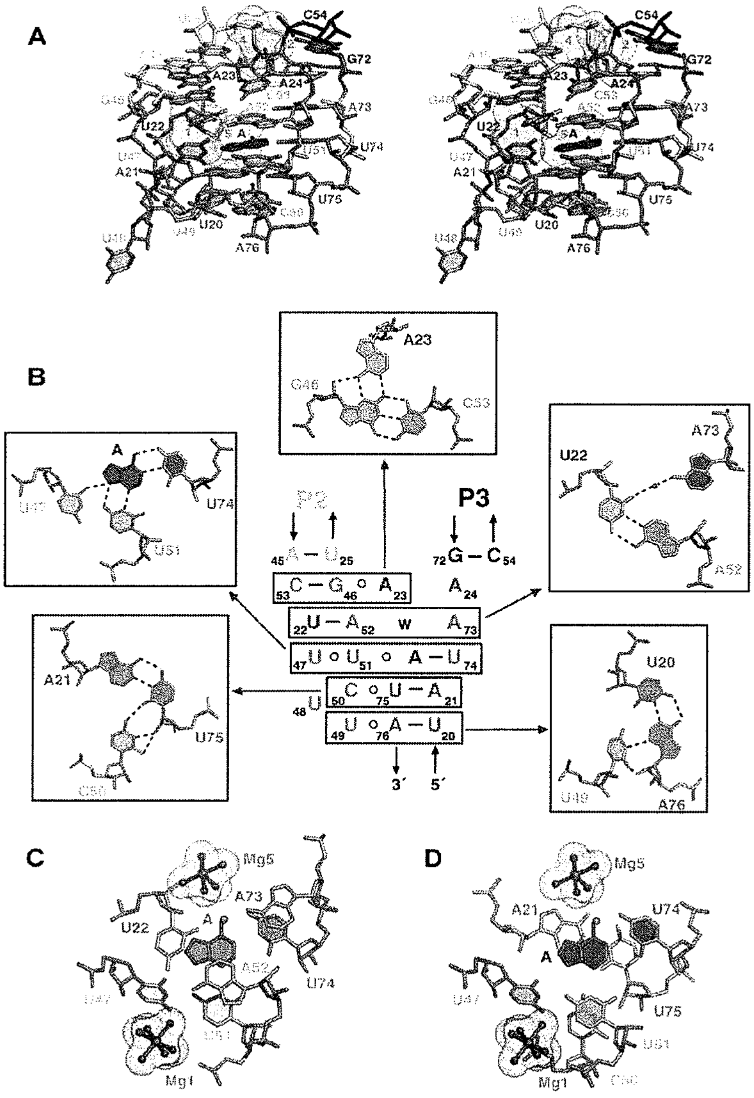


Figure 5. Details of the Tertiary Interactions within the Zippered Up Junctional Bubble in the Structure of the A-Riboswitch-Adenine Complex

(A) Stereo pair of the junction-connecting J1-2, J2-3, and J-31 segments and two junctional stem P1 base pairs that constitute the core of the complex and include the adenine binding site. The color-coding is the same as in the caption to Figure 2, with the bound adenine shown in red.

(B) Schematic of the tertiary interactions involving a five-tiered arrangement of base triples. Individual base triples are boxed, and their pairing alignments (together with assignments) are shown on either side and above the schematic. Water molecule (w) is shown as a green ball. Hydrogen bonds are indicated by dotted lines.

(C) Stacking of the A73•(A52-U22) triple over the shaded U51•(adenine-U74) recognition triple.

(D) Stacking of the shaded U51•(adenine-U74) recognition triple over the C50•(U75-A21) triple. In (C) and (D), hydrated Mg ions, surrounded by the solvent-accessible surface in a mesh representation, are shown in ball-and-stick representation.

connecting J1-2, J2-3, and J3-1 segments. Because the U51•(adenine-U74) base triple is sandwiched between the water-mediated A73•(A52-U22) and C50•(U75-A21) base triples (Figure 5A), the bound adenine is completely surrounded by RNA, both along its periphery and above and below its base plane.

Divalent Cation Binding Sites

We observe five hydrated Mg binding sites in the A-riboswitch-adenine complex, one of which is on the surface and is involved in packing interactions. The remaining four Mg sites are positioned deep within grooves, primarily involving the junction-connecting segments in the core of the fold (Figures 4B and 4C). Mg1 and Mg4 are positioned within a deep groove made up of segments of J2-3 and the beginning of helix P2, and Mg2 and Mg5 are positioned within a deep groove made up of segments of J1-2, J3-1, and the beginning of helix P3. Hydrated Mg1 and Mg5 are closest to the adenine binding pocket (Figures 5C and 5D), with Mg5 located within hydrogen bonding distance of the N^6 position of the bound adenine and thereby contributing to the locking-up of the binding pocket.

xpt G-Riboswitch-Guanine Complex Structure

Overall Topology and Tertiary Interactions

The global structure of the G-riboswitch-guanine complex is very similar to the global structure of the A-riboswitch-adenine complex (Figure 4). The same tertiary interactions hold the central core and kissing hairpins, and the purine binding pocket is sculptured by the same residue positions in both complexes. Because U62 in the A-riboswitch is replaced by G in the G-riboswitch (Figure 2), we observed formation of a G62•U63 base platform stabilized by a single hydrogen bond in the G-riboswitch.

Guanine Recognition Specificity

The bound guanine moiety in the G-riboswitch complex is held in position through the same set of hydrogen-bonding interactions observed for the bound adenine in the A-riboswitch (Figure 7), except that U74 in the

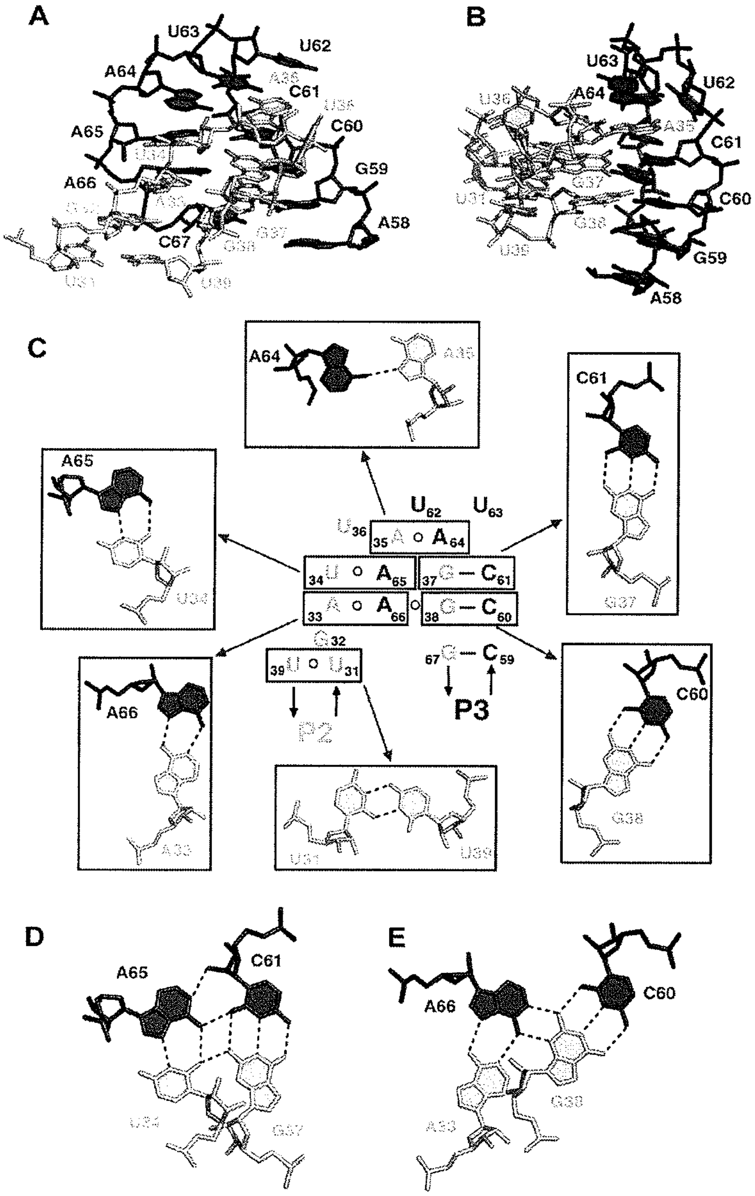


Figure 6. Details of the Tertiary Interactions between the Kissing Hairpins in the Structure of the A-Riboswitch-Adenine Complex

(A and B) Two views (rotated by 90° along the vertical axis) of the kissing interaction between loop L2, in yellow, and loop L3, in blue.

(C) Schematic of the tertiary interactions between kissing hairpin loops. Individual tertiary pairs are boxed, and their pairing alignments (together with assignments) are shown around the schematic. Pairing alignment of adjacent (D) U34•A65•C61-G37 and (E) A33•A66•C60-G38 tetrads.

A-riboswitch is replaced by C74 in the G-riboswitch. Thus, pyrimidine 74 is the specificity-determining residue in purine-sensing mRNAs. It aligns through Watson-Crick adenine-U74 pairing in the A-riboswitch and Watson-Crick guanine-C74 pairing in the G-riboswitch, as proposed earlier from mutagenesis data [24].

Discussion

Our research efforts address structural and functional issues related to folding and recognition by RNA-regulatory scaffolds associated with in vivo RNA aptamer modules. The primary challenge has been to understand the molecular recognition rules that govern ligand-mediated adaptive conformational transitions on RNA scaffolds that affect function. These principles underlie determinants of affinity and specificity, as well as global topology and allosteric transitions, thereby providing mechanistic insights into recognition processes associated with biological function.

Adaptive Conformational Transition on Complex Formation

The NMR results establish that the A-riboswitch binds adenine and G-riboswitch binds guanine selectively with 1:1 stoichiometry, with both complexes formed with high affinity (slow exchange on the NMR time scale). Complex formation occurs in the absence of divalent ions, but low (2 mM) Mg concentration was required to drive complex formation to completion.

We observe a large number of additional imino protons on proceeding from the free to the bound states of

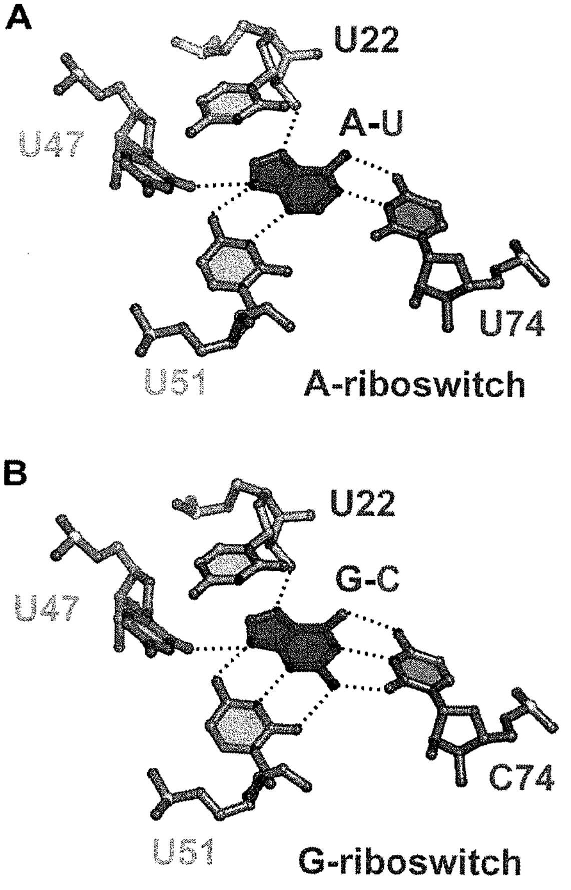


Figure 7. Recognition of Bound Purines in Purine Riboswitches
(A) Hydrogen-bonding alignments to bound adenine in the A-riboswitch. The bound adenine forms a Watson-Crick pair with U74.
(B) Hydrogen-bonding alignments to bound guanine in the G-riboswitch. The bound guanine forms a Watson-Crick pair with C74. Hydrogen bonds involving 2′-OH of U22 and base edges of U47 and U51 are common to both riboswitches. Oxygen, nitrogen, and phosphorus atoms are shown, respectively, as red, blue, and yellow balls.

the A- and G-riboswitches (Figure 3). These additional imino protons on complex formation are detected in the 13.0–14.0 ppm region characteristic of imino protons involved in N-H••N hydrogen bonds and in the 10.0–11.0 ppm region characteristic of imino protons involved in N-H••O hydrogen bonds and/or those shielded from solvent. These results are indicative of stabilization of flexible regions within the riboswitches on ligand binding associated with an adaptive conformational transition, which results in a more compact tertiary fold on complex formation. These findings are consistent with the biochemical data published previously that indicated substantial structural stabilization of the highly conserved core of the aptamers [23, 24].

Global Architecture of Purine-Riboswitch Scaffolds

Both the A-riboswitch-adenine complex and the G-riboswitch-guanine complex adopt similar global scaffolds. The tuning-fork architecture of the three helical stems is achieved by the kissing interactions between hairpin loops L2 and L3, which forces stems P2 and P3 (the prongs of the tuning fork) to align in a parallel arrangement. The tuning fork architecture of the stems in turn defines the relative positioning of the junctional Watson-Crick base pairs A21-U75 of stem P1, U25-A45 of stem P2, and C54-G72 of stem P3. These base pairs serve as bookends bracketing the base triple interactions involving residues in J1-2, J2-3, and J3-1 junction-connecting segments that form the core of the architecture and are involved in the generation of the ligand binding pocket. It is striking that both riboswitch complexes adopt the same tertiary architectures despite different crystal packing interactions, pH, and Mg crystallization conditions.

Generation of a Central Core

A remarkable feature of the ligand bound A- and G-riboswitch scaffolds is the formation of a junctional core involving five stacked triples, with the bound purine constituting an integral part of the central triple. All five triples exhibit a common theme, in which the third base is positioned in the minor groove of a Watson-Crick base pair (Figure 5B) and is aligned with either the minor groove edge and/or sugar 2′-hydroxyl group, similar to what has been reported for a RNA pseudoknot scaffold [36], the P4-P5-P6 domain of the group I intron [37], and a vitamin B_{12}-RNA aptamer complex [38]. It has been previously documented that this A-minor interaction strongly prefers Watson-Crick pairs over their noncanonical counterparts [37, 39] and that this plays a critical role in the specificity of RNA-RNA helix recognition [40].

There is a high degree of conservation of residues in the J1-2, J2-3, and J3-1 junction-connecting segments and the two junctional base pairs of stem P1 amongst the sequences of the purine-sensing mRNAs [24]. This conservation reflects the key roles these residues play in forming the stacked base triples array that defines the central core. Residue conservation is not maintained at position 24 in J1-2, position 48 in J2-3, and position 73 in J3-1, with the structures providing rational explanations for these observations. The central core architecture is anchored in place through formation of an A23-A24 platform [41], in which A23 and A24 are splayed out in the same plane, with A24 intercalating between an extended G72-A73 step. Thus, A24 acts as a spacer, while U48 loops out of the core scaffold. Residue A73 forms a water-mediated triple with the Watson-Crick A52-U22 pair (Figure 5B), and in the absence of direct interactions, the alignment is consistent with nonconservation of A at this position.

It is noteworthy that the junctional Watson-Crick A21-U75 and U20-A76 base pairs in stem P1 are highly conserved in the purine-riboswitches. This is now readily explainable because these base pairs are involved in C50•(U75-A21) and U49•(A76•U20) triple formation, with the former triple serving as the platform for stacking of the recognition triple. In contrast, the sequences of the remaining stem segments appear not to be important for riboswitch function.

Stitching Together the Ends

It was correctly anticipated that loops L2 and L3, given their complementary sequences, would pair with each other [23, 24]. Nevertheless, formation of two distinct types of *trans* A•A pairs and one *trans* A•U noncanonical pair was unexpected, as was their further alignment together with two tertiary Watson-Crick base pairs to form an adjacent pair of stacked mixed-tetrad alignments. Mixed tetrads have been reported previously in the structure of a viral pseudoknot [36] and facilitate the generation of a compact and maximally aligned kissing loop scaffold. The majority of the bases of loops L2 and L3 are involved in pairing between loops, and some of the remaining loop residues play distinct architectural roles. Hence, G32 is stacked into the scaffold as a spacer element, whereas U62 and U63 cap the interacting loop ends, and only U36 is looped out of the kissing scaffold.

The structure of the complex readily explains why the lengths of stems P2 (7 bp) and P3 (6 bp) are conserved. The observed lengths of stems P2 and P3 appear to be optimal for facilitating the observed kissing loop interactions. Such an establishment of distance and orientation constraints of stem-loop sequences P2-L2 and P3-L3 relative to the three-stem junction emphasizes the importance of peripheral tertiary kissing interactions in sculpting the topology of the ligand binding pocket, despite the large distance separating these two motifs.

Pyrimidine 74 as a Specificity Determinant

A systematic comparison between the A- and G-riboswitch sequences correctly led to the prediction that pyrimidine residue 74 is the key specificity determinant [24]. Thus, it was predicted that bound adenine would pair with U74 in the sequence of the A-riboswitch and that bound guanine would pair with C74 in the sequence of the G-riboswitch, presumably in each case through Watson-Crick pairing. Such an alignment would explain why the A-riboswitch containing U74 binds adenine and discriminates against guanine by greater than four orders of magnitude [24] and why the G-riboswitch containing C74 binds guanine and discriminates against adenine by six orders of magnitude [23]. Indeed, an A-riboswitch can be changed to a G-riboswitch simply by converting U74 to C74, and a G-riboswitch can be changed to an A-riboswitch by converting C74 to U74 [24].

Our structures of the A-riboswitch-adenine complex and the G-riboswitch-guanine complex indeed identify a Watson-Crick adenine-U74 alignment in the former complex compared to a Watson-Crick guanine-C74 alignment in the latter complex (Figure 7). These structures unambiguously demonstrate that the identity of a single nucleotide out of approximately 70 residues is sufficient for switching the specificity of a purine-sensing mRNA.

Purine Binding Pocket

Both adenine and guanine in their respective riboswitch complexes are completely surrounded by their common RNA scaffolds. Thus, constrained folding of junction-connecting segments juxtaposes structural elements to create a complementary surface for molecular recognition and discrimination against closely related analogs. Three pyrimidine rings and a sugar ring that line the binding cavity target the base edges of the bound adenine and guanine. Two of these pyrimidines, U47 and U51 from J2-3, are common to both complexes, and their Watson-Crick base edges align with the N^3-N^9H minor groove edge of both bound adenine and guanine (Figure 7). The 2′-OH group on the sugar ring of U22 from J1-2 targets the N^7 of both adenine and guanine in both complexes. The specificity determinant residue is located at pyrimidine 74, with U74 and C74 targeting the Watson-Crick edge of bound adenine and guanine, respectively. It should be noted that both amino protons of the bound guanine are hydrogen bonded, consistent with a pronounced loss of binding affinity on substitution of guanine ligand by N-methylguanine [24].

The recognition base triples involving bound adenine or guanine are sandwiched between the A73•(A52-U22) and C50•(U75-A21) base triples (Figure 5A). The bound adenine or guanine was found to stack over both flanking triples (Figures 5C and Figures 5D). Such stacking interactions of planar purines sandwiched between base triple platforms have been previously reported for in vitro selected RNA aptamer complexes with bound planar ligands.

The bound purine ring systems are surrounded by RNA along their periphery as well as above and below their planar ring systems. Such an encapsulated recognition architecture readily explains why every functionalized position on the adenine/guanine heterocycle, even for closely related analogs, results in a substantial loss in binding affinity [23, 24]. The dramatic decrease in the majority of the analog binding affinities can be attributed to disruption of intermolecular hydrogen bonding interactions, and the remainder can be attributed to the consequences of steric occlusion.

Comparison of Adenine, Guanine, and Hypoxanthine Bound Riboswitches

The crystal structure of the hypoxanthine bound riboswitch [33] and our crystal structures of the adenine bound and guanine bound riboswitches are in excellent agreement. The crystallographic phases of the hypoxanthine complex were obtained through anomalous scattering of $Co(NH_3)_6$, whereas the phases in our structure of the adenine complex were solved through anomalous scattering properties of $BaCl_2$. Hypoxanthine lacks the 2-amino group present in guanine, and, hence, the crystal structure of the hypoxanthine bound G-riboswitch [33] differs from its guanine bound G-riboswitch counterpart in the absence of two hydrogen bonds associated with 2-amino group recognition. The uniqueness of our contribution comparing the structures of bound A- and G-riboswitches reflects our ability to definitively establish the recognition principles by which a single specificity-determining residue permits discrimination between closely related metabolites.

Divalent Cation-Mediated Stabilization

There is ample documentation in the RNA literature that Mg cations can mediate interactions between structural domains, especially in regions where different segments of the phosphodiester backbone are brought in close proximity [42–44]. Such a stabilization effect of bound Mg cations is also observed in the crystal structure of the A-riboswitch-adenine complex (Figures 4B and Figures 4C) for crystals grown under high (200 mM) Mg concentration. In this structure, four Mg cations are positioned deep in grooves formed by junction-connecting segments in the core of the complex.

We have not been able to detect any bound hydrated Mg cations in the crystal structure of the G-riboswitch-guanine complex for crystals grown under lower (20 mM) Mg concentration. Our imino proton NMR spectra indicate that complex formation can occur in the absence of added Mg, but adding 2 mM Mg drives complex formation to completion.

Significance

Riboswitches are an important form of genetic control in certain bacteria, where they modulate the expression of numerous metabolic and transport proteins. Each riboswitch class carries a distinctive sequence and structure that is highly conserved among the organisms that use this form of regulatory system. With X-ray crystallography, we have determined high-resolution structures of the related purine-specific riboswitches that selectively bind adenine or guanine, complementing a parallel structure determination of a hypoxanthine bound riboswitch [33]. Although the A- and G-riboswitches only share 60% sequence identity, they form nearly identical binding pockets for their corresponding ligands. A single nucleotide within the core forms a Watson-Crick base pair with the ligand and, thus, serves as the main selectivity determinant between these two natural ligands. The structures of the A- and G-riboswitches bound to their respective ligands establish RNA's ability to utilize A-minor motifs and base tetrads to facilitate folding of junctional architectures and thereby generate occluded binding pockets. Such scaffolds provide the infrastructure for targeting and discrimination between closely related metabolites, on the basis of precise hydrogen bonding and shape complementarity. In addition, the kissing interaction between the hairpin loops in the purine-riboswitch appears to be critical for both global scaffold formation and binding pocket architecture and thereby mediates long-range effects on ligand binding and release.

Experimental Procedures

RNA Preparation

The DNA fragments for riboswitch production were prepared by the annealing of chemically synthesized oligonucleotides. The DNA fragments were placed under control of the T7 promoter by cloning into StuI and HindIII sites of the pUT7 vector [45]. The *add* and *xpt* mRNAs were synthesized by in vitro transcription with T7 RNA polymerase with linearized plasmid DNA as templates. Self-cleaving hammerhead ribozyme sequence was added at the 3′-end of each RNA for subsequent cleavage to ensure length homogeneity. Purification of the transcripts was performed on 15% denaturing gels and was followed by cation exchange chromatography on MonoQ column (Amersham) and ethanol precipitation.

Complex Formation and Crystallization

The *add* A-riboswitch-adenine complex was prepared at 0.7 mM concentration in 50 mM potassium acetate buffer (pH 6.8) in the presence of 2 mM $MgCl_2$. Because of the low solubility of guanine, the *xpt* G-riboswitch-guanine complex was first prepared at 10 μM concentration and then lyophilized to reduce the volume 100-fold. Each sample was combined with the same volume of the reservoir solution, and crystals were grown by the hanging-drop vapor diffusion method. Reservoir solutions were: (1) A-riboswitch-adenine complex, 3 M 1,6-hexanediol, 0.1 M Tris-HCl (pH 9.0), and 200 mM $MgCl_2$; (2) G-riboswitch-guanine complex, 28% PEG400, 100 mM sodium citrate, 300 mM ammonium acetate (pH 5.2), 20 mM $MgCl_2$, and 1% spermine. Crystals of the complexes grew to a maximal size of 300 × 50 × 50 μm in approximately 7 days at +4°C.

Structure Determination

The barium derivative was obtained by soaking the A-riboswitch-adenine complex crystals in stabilizing solution supplemented with 60 mM $BaCl_2$ for 3 days. Crystals were flash-frozen in liquid nitrogen. The 2.7 Å resolution native and 2.95 Å resolution Ba-derivative data sets for the A-riboswitch-adenine complex were collected on the in-house Rigaku diffractometer at 1.54 Å wavelength. Other X-ray data (2.1 Å A-riboswitch-adenine complex and 2.4 Å G-riboswitch-guanine complex) were collected at beamline X25 at the National Synchrotron Light Source (NSLS). Data were processed with HKL2000 (HKL Research, Charlottesville, VA).

To determine the structure of the A-riboswitch-adenine complex, we first located six barium sites by single isomorphous replacement and anomalous scattering technique (SIRAS). The sites were found with SHELXD software [46] with the isomorphous 2.95 Å derivative and 2.7 Å native data. The position of the sites was refined with Mlphare [47]. The SIRAS phasing was performed with SHARP [48] and included the density modification procedure with SOLOMON [49] assuming the optimized solvent content 40%. The initial model was built manually with the SIRAS electron density map and program O [50]. The model was then refined for the resolution range 20.0–2.1 Å with the higher resolution 2.1 Å native data set, as processed by CNS [51] and REFMAC [52] programs. Electron density was of sufficient quality to unambiguously interpret the entire structure except for the 5′-triphosphate and 3′-cyclic phosphate. Data collection, phasing, and refinement statistics are listed in Table S1. Adenine, hydrated Mg^{2+} cations, and water molecules were added to the model at the final stage of the refinement on the basis of analysis of the 2Fo-Fc and the difference Fo-Fc electron density maps.

The structure of the G-riboswitch-guanine complex was determined by molecular replacement for the resolution range 10.0–4.0 Å with Molrep software [47] with the A-riboswitch-adenine complex structure as a search model. The model was refined with the 2.4 Å resolution data set with protocols similar to those used to solve the A-riboswitch-adenine complex structure. The model contains all residues including 5′-triphosphate and 3′-cyclic phosphate.

NMR Experiments

Imino proton NMR spectra were recorded with Varian and Bruker NMR spectrometers at 25°C, with jump-and-return (JR) water suppression for detection [53]. The NMR sample conditions are listed in the caption to Figure 3.

Graphics

The figures were prepared with PyMOL (http://pymol.sourceforge.net/) and nuccyl (http://www.mssm.edu/students/jovinl02/research/nuccyl.html) software.

Supplemental Data

Crystallographic data for the riboswitch-ligand complexes, as well as NMR spectra for the *ydhL* A-riboswitch bound to various ligands, are available at http://www.chembiol.com/cgi/content/full/11/12/1729/DC1/.

Acknowledgments

This research was supported by National Institutes of Health grant GM73618, the DeWitt Wallace Foundation, and the Abby Rockefeller Mauze Trust to D.J.P., National Institutes of Health grant GM068819 and National Science Foundation grant EIA-0323510 to R.R.B., and the Austrian Science Fund (FWF) to R.M. We thank P.

Saha for initial experiments, A. Teplov for help with data collection, and M. Becker and the staff of the X25 and X12C beam lines at the National Synchrotron Light Source for assistance with data collection. D.J.P. is a member of the New York Structural Biology Center, which is supported by National Institutes of Health grant GM66354.

Received: November 24, 2004
Revised: November 30, 2004
Accepted: November 30, 2004
Published: December 20, 2004

References

1. Batey, R.T., Rambo, R.P., and Doudna, J.A. (1999). Tertiary motifs in RNA structure and folding. Angew. Chem. Int. Ed. Engl. *38*, 2326–2343.
2. Hermann, T., and Patel, D.J. (1999). Stitching together RNA tertiary architectures. J. Mol. Biol. *294*, 829–849.
3. Doherty, E.A., and Doudna, J.A. (2001). Ribozyme structure and mechanisms. Annu. Rev. Biophys. Biomol. Struct. *30*, 457–475.
4. Lilley, D.M.J. (1999). Structure, folding and catalysis of the small nucleolytic ribozymes. Curr. Opin. Struct. Biol. *9*, 330–338.
5. Jaschke, A. (2001). Artificial ribozymes and deoxyribozymes. Curr. Opin. Struct. Biol. *11*, 321–326.
6. Nowakowski, J., and Tinoco, I., Jr. (1999). RNA structure in solution. In Oxford Handbook of Nucleic Acid Structure, S. Neidle, ed. (New York: Oxford University Press), pp. 567–602.
7. Patel, D.J. (1999). Adaptive recognition in RNA complexes with peptide and protein molecules. Curr. Opin. Struct. Biol. *9*, 74–87.
8. Leulliot, N., and Varani, G. (2001). Current topics in RNA-protein recognition: Control of specificity and biological function through induced fit and conformational capture. Biochemistry *40*, 7947–7956.
9. Famulok, M., Mayer, G., and Blind, M. (2000). Nucleic acid aptamers: From selection *in vitro* to applications *in vivo*. Acc. Chem. Res. *33*, 591–599.
10. Hermann, T., and Patel, D.J. (2000). Adaptive recognition by nucleic acid aptamers. Science *287*, 820–825.
11. Soukup, G.A., and Breaker, R.R. (1999). Nucleic acid molecular switches. Trends Biotechnol. *17*, 469–476.
12. Soukup, G.A., and Breaker, R.R. (2000). Allosteric nucleic acid catalysts. Curr. Opin. Struct. Biol. *10*, 318–325.
13. Breaker, R.R. (2002). Engineered allosteric ribozymes as biosensor components. Curr. Opin. Biotechnol. *13*, 31–39.
14. Soukup, G.A., and Breaker, R.R. (1999). Engineering precision molecular switches. Proc. Natl. Acad. Sci. USA *96*, 3584–3589.
15. Kaempfer, R. (2003). RNA sensors: Novel regulators for gene expression. EMBO Rep. *4*, 1043–1047.
16. Nahvi, A., Sudarsan, N., Ebert, M.S., Zou, X., Brown, K.L., and Breaker, R.R. (2002). Genetic control by a metabolite-binding RNA. Chem. Biol. *9*, 1043–1049.
17. Mironov, A.S., Gusarov, I., Rafikov, R., Lopez, L.E., Shatalin, K., Kreneva, R.A., Perumov, D.A., and Nudler, E. (2002). Sensing small molecules by nascent RNA: A mechanism to control transcription in bacteria. Cell *111*, 747–756.
18. Winkler, W., Cohen-Chalamish, S., and Breaker, R.R. (2002). An mRNA structure that controls gene expression by binding FMN. Proc. Natl. Acad. Sci. USA *99*, 15908–15913.
19. Winkler, W., Nahvi, A., and Breaker, R.R. (2002). Thiamine derivatives bind mRNAs directly to regulate bacterial gene expression. Nature *419*, 952–956.
20. McDaniel, B.A.M., Grundy, F.J., Artsimovitch, I., and Henkin, T.M. (2003). Transcription termination control of the S box system: Direct measurement of *S*-adenosylmethionine by the leader RNA. Proc. Natl. Acad. Sci. USA *100*, 3083–3088.
21. Epshtein, V., Mironov, A.S., and Nudler, E. (2003). The riboswitch-mediated control of sulfur metabolism in bacteria. Proc. Natl. Acad. Sci. USA *100*, 5052–5056.
22. Winkler, W.C., Nahvi, A., Sudarsan, N., Barrick, J.E., and Breaker, R.R. (2003). An mRNA structure that controls gene expression by binding *S*-adenosylmethionine. Nat. Struct. Biol. *10*, 701–707.
23. Mandal, M., Boese, B., Barrick, J.E., Winkler, W.C., and Breaker, R.R. (2003). Riboswitches control fundamental biochemical pathways in *Bacillus subtilisis* and other bacteria. Cell *113*, 577–586.
24. Mandal, M., and Breaker, R.R. (2004). Adenine riboswitches and gene activation by disruption of a transcription terminator. Nat. Struct. Mol. Biol. *11*, 29–35.
25. Grundy, F.J., Lehman, S.C., and Henkin, T.M. (2003). The L box regulon: Lysine sensing by leader RNAs of bacterial lysine biosynthesis genes. Proc. Natl. Acad. Sci. USA *100*, 12057–12062.
26. Sudarsan, N., Wickiser, J.K., Nakamura, S., Ebert, M.S., and Breaker, R.R. (2003). An mRNA structure in bacteria that controls gene expression by binding lysine. Genes Dev. *17*, 2688–2697.
27. Mandal, M., Lee, M., Barrick, J.E., Weinberg, Z., Emilsson, G.M., Ruzzo, W.L., and Breaker, R.R. (2004). A glycine-depedent riboswitch that uses cooperative binding to control gene expression. Science *306*, 275–279.
28. Nudler, E., and Mironov, A.S. (2004). The riboswitch control of bacterial metabolism. Trends Biochem. Sci. *29*, 11–17.
29. Mandal, M., and Breaker, R.R. (2004). Gene regulation by riboswitches. Nat. Rev. Mol. Cell Biol. *5*, 451–463.
30. Winkler, W.C., Nahvi, A., Roth, A., Collins, J.A., and Breaker, R.R. (2004). Control of gene expression by a natural metabolite-responsive ribozyme. Nature *428*, 281–286.
31. Barrick, J.E., Corbino, K.A., Winkler, W.C., Nahvi, A., Mandal, M., Collins, J., Lee, M., Roth, A., Sundarsan, N., Jona, I., et al. (2004). New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. Proc. Natl. Acad. Sci. USA *101*, 6421–6426.
32. Sudarsan, N., Barrick, J.E., and Breaker, R.R. (2003). Metabolite-binding RNA domains are present in the genes of eukaryotes. RNA *9*, 644–647.
33. Batey, R.B., Gilbert, S.D., and Montagne, R.K. (2004). Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine. Nature *432*, 411–415.
34. Leontis, N.B., Stombaugh, J., and Westhof, E. (2002). The non-Watson-Crick base pairs and their associated isostericity matrices. Nucleic Acids Res. *30*, 3497–3531.
35. Leontis, N.B., and Westhof, E. (2003). Analysis of RNA motifs. Curr. Opin. Struct. Biol. *13*, 300–308.
36. Su, L., Chen, L., Egli, M., Berger, J.M., and Rich, A. (1999). Minor groove RNA triplex in the crystal structure of a ribosomal frameshifting viral pseudoknot. Nat. Struct. Biol. *6*, 285–292.
37. Doherty, E.A., Batey, R.T., Masquida, B., and Doudna, J.A. (2001). A universal mode of helix packing in RNA. Nat. Struct. Biol. *8*, 339–343.
38. Sussman, D., Nix, J.C., and Wilson, C. (2000). The structural basis for molecular recognition by the vitamin B_{12} RNA aptamer. Nat. Struct. Biol. *7*, 53–57.
39. Nissen, P., Ippolito, J.A., Ban, N., Moore, P.B., and Steitz, T.A. (2001). RNA tertiary interactions in the large ribosomal subunit: The A-minor motif. Proc. Natl. Acad. Sci. USA *98*, 4899–4903.
40. Battle, D.J., and Doudna, J.A. (2002). Specificity of RNA-RNA helix recognition. Proc. Natl. Acad. Sci. USA *99*, 11676–11681.
41. Cate, J.H., Gooding, A.R., Podell, E., Zhou, K., Golden, B.L., Szewczak, A.A., Kundrot, C.E., Cech, T.R., and Doudna, J.A. (1996). RNA tertiary structure mediation by adenosine platforms. Science *20*, 1696–1699.
42. Hanna, R., and Doudna, J.A. (2000). Metal ions in ribozyme folding and catalysis. Curr. Opin. Chem. Biol. *4*, 166–170.
43. Egli, M., Minasov, G., Su, L., and Rich, A. (2002). Metal ions and flexibility in a viral pseudoknot at atomic resolution. Proc. Natl. Acad. Sci. USA *99*, 4302–4307.
44. Klein, D.J., Moore, P.B., and Steitz, T.A. (2004). The contribution of metal ions to the structural stability of the large ribosomal subunit. RNA *10*, 1366–1379.
45. Serganov, A., Rak, A., Garber, M., Reinbolt, J., Ehresmann, B., Ehresmann, C., Grunberg-Manago, M., and Portier, C. (1997). Ribosomal protein S15 from *Thermus thermophilus*. Cloning, sequencing, overexpression of the gene and RNA-binding properties of the protein. Eur. J. Biochem. *246*, 291–300.

46. Schneider, T.R., and Sheldrick, G.M. (2002). Substructure solution with SHELXD. Acta Crystallogr. D Biol. Crystallogr. *58*, 1772–1779.
47. CCP4 (Collaborative Computational Project, Number 4)(1994). The CCP4 suite: programs for protein crystallography. Acta Crystallogr. D Biol. Crystallogr. *50*, 760–763.
48. De La Fortelle, E., and Bricogne, G. (1997). Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous and multiwavelength anomalous diffraction methods. Methods Enzymol. *276*, 472–494.
49. Abrahams, J.P., and Leslie, A.G.W. (1996). Methods used in the structure determination of bovine mitochondrial F1 ATPase. Acta Crystallogr. D Biol. Crystallogr. *52*, 30–42.
50. Jones, T.A., Zou, J.Y., Cowan, S.W., and Kjeldgaard, G.J. (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Crystallogr. A *47*, 110–119.
51. Brunger, A.T., Adams, P.D., Clore, G.M., De Lano, W.L., Gros, P., Grosse-Kuntsleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., et al. (1998). Crystallography and NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr. D Biol. Crystallogr. *5*, 905–921.
52. Murshudov, G.N., Vagin, A.A., and Dodson, E.J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr. D Biol. Crystallogr. *53*, 240–245.
53. Phan, A.T., Gueron, M., and Leroy, J.L. (2001). Investigation of unusual DNA motifs. Methods Enzymol. *338*, 341–371.

Accession Numbers

Coordinates for the 2.1 Å complex of adenine bound to the adenine-sensing *add* mRNA and the 2.4 Å complex of guanine bound to the guanine-sensing *xpt* mRNA have been deposited in the Protein Data Bank under accession codes 1Y26 and 1Y27, respectively.

Full text provided by www.sciencedirect.com

Riboswitches as versatile gene control elements

Brian J Tucker[1] and Ronald R Breaker[2]

Riboswitches are structured elements typically found in the 5′ untranslated regions of mRNAs, where they regulate gene expression by binding to small metabolites. In all examples studied to date, these RNA control elements do not require the involvement of protein factors for metabolite binding. Riboswitches appear to be pervasive in eubacteria, suggesting that this form of regulation is an important mechanism by which metabolic genes are controlled. Recently discovered riboswitch classes have surprisingly complex mechanisms for regulating gene expression and new high-resolution structural models of these RNAs provide insight into the molecular details of metabolite recognition by natural RNA aptamers.

Addresses
[1] Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA
[2] Department of Molecular, Cellular and Developmental Biology, Yale University, PO Box 208103, New Haven, CT 06520, USA

Corresponding author: Breaker, Ronald R (ronald.breaker@yale.edu)

Current Opinion in Structural Biology 2005, **15**:342–348

This review comes from a themed issue on
Nucleic acids
Edited by Dinshaw J Patel and Eric Westhof

Available online 24th May 2005

0959-440X/$ – see front matter

DOI 10.1016/j.sbi.2005.05.003

Introduction

From microRNAs [1] to 'riboregulators' [2], one of the more salient concepts to have emerged from gene regulation research over the past several years is that RNA frequently plays a more direct and intimate role in controlling gene expression than previously assumed [3,4]. It has long been known that differential folding of RNA plays a major role in transcriptional attenuation [5]. Other RNA-based regulatory mechanisms have subsequently been discovered, including pathways involving antisense [6] and tRNA–mRNA interactions [7], control of translation by temperature-dependent modulation of RNA structure [8–11] and the involvement of microRNAs as *trans*-acting genetic factors [1]. The frequency at which these discoveries have been occurring suggests an even greater role for RNA in cellular control processes. This already appears to be true of bacteria, as recent descriptions of gene control by riboswitches are revealing a pervasive system of RNA-mediated gene control [12–16].

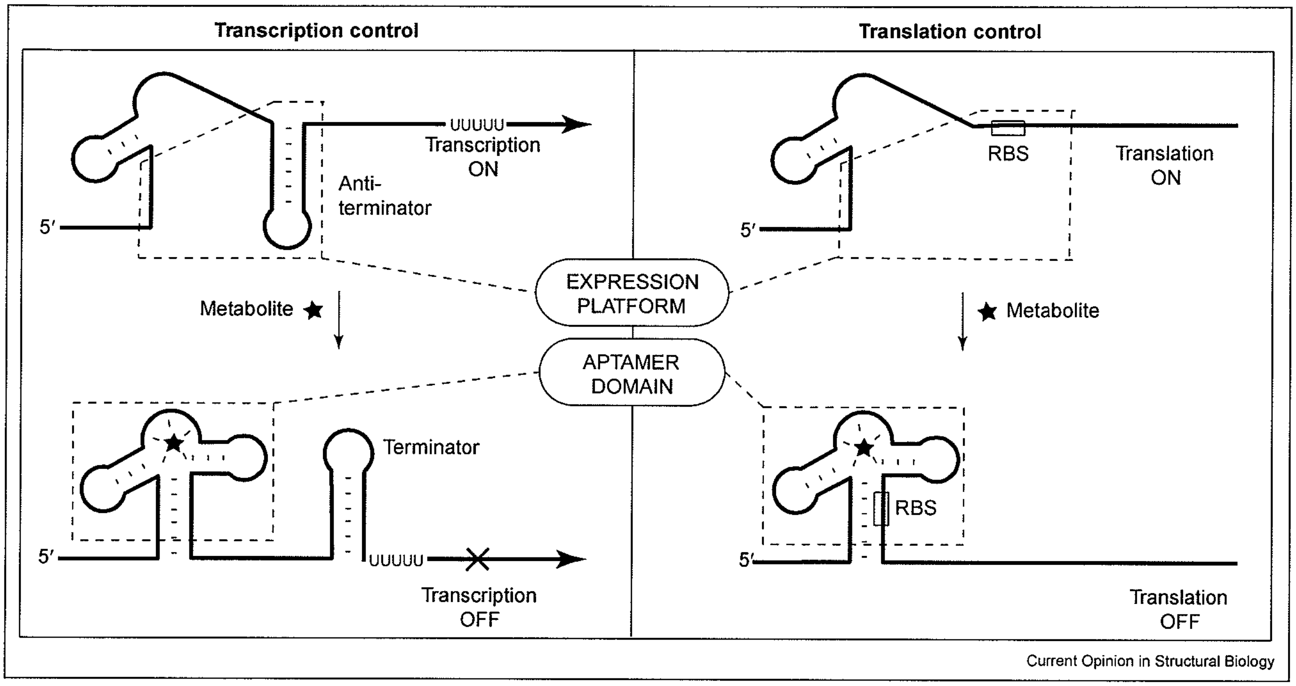
Riboswitches are widespread in bacteria, with nine classes already reported, some of which describe previously known conserved regulatory elements (e.g. [17–19]). The metabolites sensed by riboswitches are diverse, among them guanine [20], flavin mononucleotide (FMN) [21,22] and lysine [23•–25•]. One riboswitch class that binds thiamine pyrophosphate (TPP) [26] also has been found in plants and fungi [27•]. However, riboswitches that occur in eubacteria have been most extensively studied and this review will primarily focus on recent advances concerning bacterial representatives. For a more comprehensive overview of this research area, the reader is directed to several recent reviews [13–15,28–30].

Structural and functional domains of riboswitches

Most riboswitches can be divided roughly into two structural domains: an aptamer [31,32] and an expression platform (Figure 1) [26]. The aptamer domain is a highly folded structure that selectively binds to the target metabolite. The expression platform converts metabolite binding events into changes in gene expression by harnessing changes in RNA folding that are brought about by ligand binding. It is the distinctive sequence and structural features of each aptamer that are used to classify each new type of riboswitch. These metabolite-binding domains are conserved amongst widely divergent organisms, indicating that they have long persisted through evolution despite the fact that protein factors should have been strong competition for billions of years. In contrast, the sequences and structures that comprise each expression platform vary considerably among different riboswitch classes and also among representatives of the same riboswitch class, even when the regulatory mechanisms employed are similar [25•,26,33].

The genes controlled by riboswitches often encode proteins involved in the biosynthesis or transport of the metabolite being sensed [30]. Therefore, the riboswitch is used as a form of feedback inhibition, whereby binding of the metabolite to the riboswitch decreases the expression of the gene products used to make the metabolite. In most instances, repression is accomplished either by terminating transcription to prevent the production of full-length mRNAs or by preventing translation initiation once a full-length mRNA has been made. For transcription control, an anti-terminator structure is formed in the absence of a bound metabolite. When a metabolite binds, a competing stem structure is formed that serves as an intrinsic transcription terminator [22,34,35]. For control of translation initiation, the bound aptamer precludes translation by rendering inaccessible the ribosome-

Figure 1



Common mechanisms of riboswitch gene control. Transcription control involves metabolite binding and stabilization of a specific conformation of the aptamer domain that precludes formation of a competing anti-terminator stem. This allows formation of a terminator stem, which prevents the full-length mRNA from being synthesized. In contrast, control of translation is accomplished by metabolite-induced structural changes that sequester the ribosome-binding site (RBS), thereby preventing the ribosome from binding to the mRNA.

binding site or Shine–Dalgarno sequence (Figure 1) [26]. This mechanism is consistent with the observation that ribosomes do not bind to an mRNA carrying a coenzyme B_{12} riboswitch that controls gene expression at the level of translation [18].

The above descriptions of riboswitch architecture and function had represented the extent of what was known about the gene control mechanisms of riboswitches. However, research over the past year has revealed new riboswitch classes and some new mechanisms by which metabolite binding leads to changes in gene expression. Newly discovered riboswitches have deviated, in some cases surprisingly, from the more common regulatory mechanisms described previously. Moreover, new X-ray crystal structures have given us a first glimpse at the molecular details of riboswitches that, until recently, could only be inferred from biochemical data.

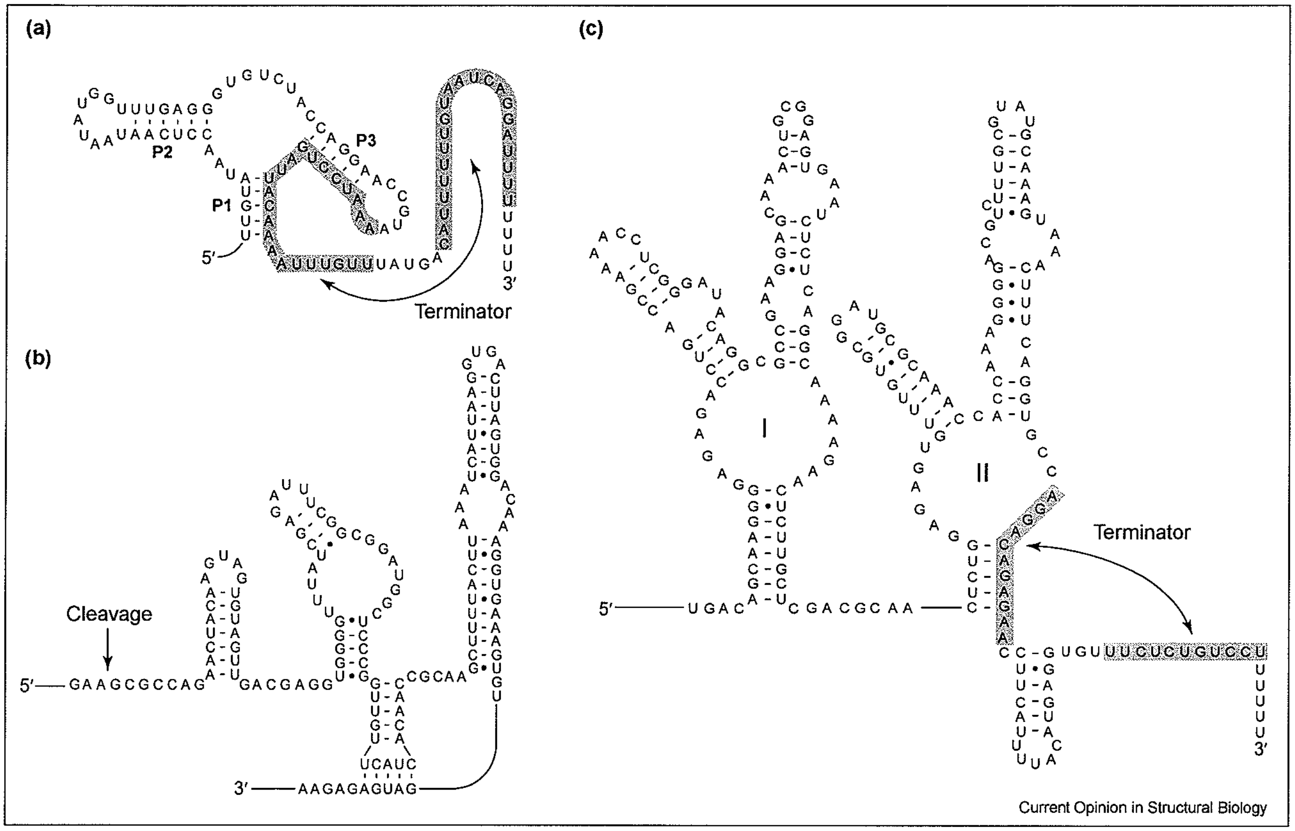
New riboswitch classes

In the past year, three newly confirmed riboswitch classes have been reported (Figure 2). In all three cases, differences are observed compared to the prototypic riboswitch mechanisms described above. The first of these, an adenine-sensing riboswitch, is remarkably similar to a previously reported riboswitch that binds guanine [20,36•,37]. The guanine riboswitch had been shown to repress purine biosynthetic and salvage genes upon binding directly to guanine [20,37]. However, certain RNA elements that conform almost perfectly to the consensus sequence and structure of guanine riboswitches were found in front of genes that were defined as coding for adenine deaminase enzymes or identified as encoding a purine efflux pump [37]. Additionally, these representative RNA elements were distinguished by the presence of a uracil residue in place of an otherwise strictly conserved cytosine residue.

This led to speculation that this single base change might alter ligand specificity. Specifically, if the cytosine (corresponding to position 74 in constructs described below) forms a Watson–Crick base pair with guanine, then mutation to a uracil would change the ligand specificity to adenine. As predicted, representative riboswitches that carry this C→U mutation are uniquely specific for adenine [36•,37].

Furthermore, these riboswitches activate the expression of a reporter gene *in vivo* upon the addition of adenine to cultured bacterial cells [36•]. In this case, it is believed that the transcription terminator stem is allowed to form only when the ligand-receptive fold of the aptamer is not

Figure 2



Current Opinion in Structural Biology

Recently characterized riboswitch classes. **(a)** Adenine riboswitch (*ydhL*). In contrast to previously characterized riboswitches, binding of adenine promotes mRNA transcription by preventing formation of a terminator stem. **(b)** GlcN6P riboswitch (*glmS*). Binding of GlcN6P induces the riboswitch to self-cleave (site identified by arrow). RNA cleavage results in decreased gene expression through an unknown mechanism. **(c)** Glycine riboswitch (*gcvT*). Binding of glycine allows mRNA transcription using a mechanism similar to that of the adenine riboswitch. However, two glycine molecules are bound cooperatively to two aptamer domains (labeled I and II). The sequence that forms the terminator stem is shaded.

formed (Figure 2a). This example of gene activation upon riboswitch–metabolite complex formation is quite rare. With all previously studied riboswitch mechanisms that control transcription termination, the terminator stem forms only when the aptamer is stabilized by metabolite binding [12]. Regardless of whether a riboswitch activates or deactivates gene expression upon ligand binding, they harness the same types of RNA folding changes. Therefore, the strong bias in favor of genetic ‘OFF’ switches is not due to any inherent limitations of RNA. Rather, this distribution of mechanisms most likely reflects the greater need that cells have for repressing metabolic gene expression when specific metabolites are in abundance.

Additional riboswitch classes have been identified by a bioinformatics search using intergenic sequences from the genome of *Bacillus subtilis* [38]. The sequence of each intergenic region (IGR) from *B. subtilis* was compared to the IGR sequences of 90 other bacteria to identify highly conserved RNA elements. Eight RNA motifs with at least some characteristics that are suggestive of riboswitch function were identified. Of these candidates, two have subsequently been shown to be riboswitches. One of these riboswitches always occurs adjacent to the *glmS* gene, which encodes the enzyme glutamine-frucotse-6-phosphate amidotransferase. This enzyme produces glucosamine-6-phosphate (GlcN6P) and it is this compound that triggers riboswitch function. Interestingly, members of the GlcN6P class of riboswitch are also self-cleaving ribozymes that are activated when the sugar-phosphate compound is bound (Figure 2b) [39••]. Mutations that diminish or abolish ribozyme self-cleavage activity *in vitro* similarly cause a reduction or loss of gene regulation *in vivo*, which suggests that ribozyme self-cleavage activity is necessary for down-regulation of gene expression. The site of RNA cleavage is located upstream of the *glmS* open reading frame. Therefore it is not clear how cleavage activity causes a reduction of gene expression. One possible explanation is that the truncated mRNA produced by the cleavage event is subsequently targeted for degradation by RNases.

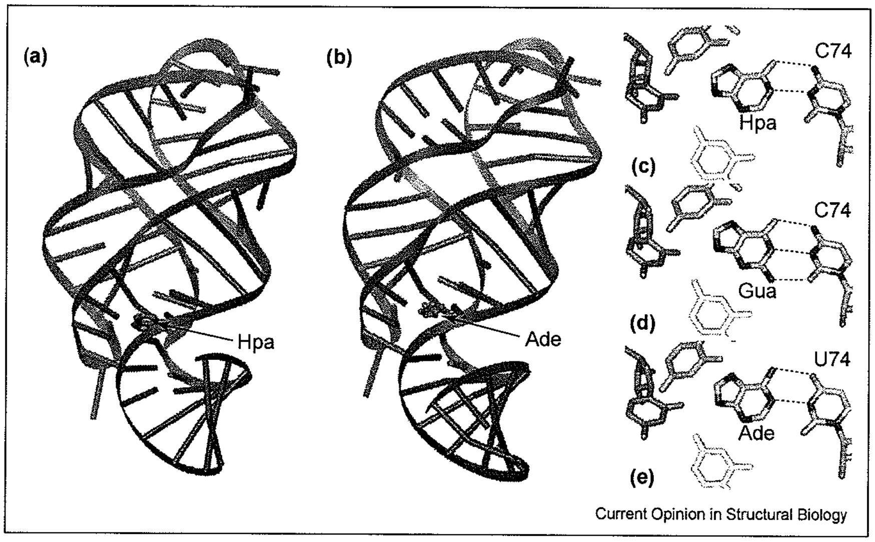
Another riboswitch candidate identified by the bioinformatics effort described above has recently been shown to function as a glycine-dependent riboswitch [40••]. Glycine riboswitches are unique in that they possess two similar aptamer structures that reside adjacent to each other, separated only by a short conserved linker sequence (Figure 2c). In many instances, these structures are located upstream of genes encoding proteins that form the glycine cleavage system, which catalyzes the initial reactions for the use of glycine as an energy source. It was shown both *in vitro* and *in vivo* that the riboswitch from *B. subtilis* activates transcription upon exposure to glycine [40••]. Thus, like the adenine riboswitch described above, members of this class serve as a rare form of genetic 'ON' switch.

The glycine riboswitch is remarkable for two other reasons. First, most glycine riboswitches carry two aptamers that each sense a ligand that has only ten atoms. Second, the two aptamers bind glycine cooperatively, such that ligand binding by one aptamer improves the binding affinity of the other by ~1000-fold and vice versa. This characteristic was demonstrated by conducting binding assays and *in vitro* transcription termination assays using constructs based on the riboswitch from *Vibrio cholerae* carrying either single or tandem aptamer configurations. For example, single and tandem aptamer constructs bind glycine with Hill coefficients of 0.97 and 1.64, respectively [40••]. Hill coefficients represent degrees of cooperativity; values near 1 represent little or no cooperativity, whereas values greater than 1 represent positive cooperativity [41]. The Hill coefficient of 1.64 supports the hypothesis that the tandem aptamer construct binds two glycine molecules cooperatively to serve as a more digital genetic switch. Furthermore, the level of cooperativity exhibited by the two aptamers of this riboswitch is comparable to that exhibited by each domain of the tetrameric protein subunits of hemoglobin [42]. Glycine riboswitches are far more sensitive to small changes in metabolite concentration than riboswitches that carry single aptamers. This more sophisticated riboswitch design is most probably required by the cell to maximize the expression of glycine cleavage system proteins when excess glycine is present. Likewise, this cooperative riboswitch can rapidly turn off the expression of these proteins as glycine levels modestly decline to assure sufficient amounts of this amino acid are available for protein synthesis.

High-resolution riboswitch structures

Recently, three high-resolution structural models have detailed the structural basis of ligand recognition by purine-binding riboswitches (Figure 3a,b) [43••,44••]. One of these structures is the aptamer domain of the guanine riboswitch from the *xpt-pbuX* operon of *B. subtilis* [43••] (Figure 3a). The aptamer was crystallized in

Figure 3



Structural features of the purine-responsive riboswitch aptamers. Structures of **(a)** the guanine riboswitch aptamer (*xpt-pbuX*) and **(b)** the adenine riboswitch aptamer (*add*) reveal a similar tertiary fold, despite only 59% sequence identity. Bound metabolites, hypoxanthine (Hpa) and adenine (Ade), respectively, are shown in red. Discrimination at the binding site of the guanine aptamer **(c,d)** and the adenine aptamer **(e)** results from the identity of nucleotide 74 (using the numbering system of the *B. subtilis* construct), which makes Watson–Crick hydrogen bonds to the appropriate metabolite. Nucleotides U22, U47 and U51 are colored orange, magenta and yellow, respectively. The oxygen and nitrogen atoms of the metabolite and nucleotide 74 are colored red and blue, respectively.

complex with hypoxanthine, a metabolite that is similar in structure to guanine, and has previously been shown to be bound by the riboswitch and to regulate gene expression [20,37,45]. In a similar study [44••], the same riboswitch was crystallized in complex with guanine.

These structural models add new insight into how RNA molecules, either natural or engineered, selectively bind to nucleotide-like compounds [46–49]. Similarities exist between this model and the structures of engineered aptamers, particularly in the use of base triples to form the binding pocket and in the use of base stacking to stabilize the bound aptamer state [43••,44••,49]. A unique feature of the riboswitch structure is a novel loop–loop interaction that stably bridges two stems of the aptamer. Mutations of these loop sequences were known to eliminate metabolite binding *in vitro* and this effect probably explains why the identities of these nucleotides are strictly conserved in all examples identified to date [20]. This loop–loop interaction involves the formation of a web of hydrogen bonds that appears to be essential for constraining the tertiary structure of the ligand-bound state [43••,44••,48] (Figure 3a).

Another notable feature of this structure is that the ligand is almost entirely engulfed by the binding pocket of the aptamer [43••,44••]. The side walls of the binding pocket or 'compartment' are formed by three nucleobases and a ribose (Figure 3c,d), whose functional groups form as many as eight hydrogen bonds with guanine. This extensive network of contacts would be difficult to achieve with most engineered aptamers, which have more open binding pockets, and typically exhibit weaker and less selective interactions with their target ligands.

As described above, most riboswitches must alternate between two structural states: one that allows gene expression and one that precludes expression (Figure 1). The ligand-bound state of the guanine aptamer allows a terminator stem to form by preventing antiterminator formation [20]. Therefore, the formation of the aptamer structure that can receive guanine must be somewhat transitory to allow these alternate structures to form. The enclosed binding compartment of the structure prevents the aptamer from completely forming in the absence of metabolite, otherwise the compartment would not allow ligand access. Furthermore, it appears that the base contributes to the folding of the aptamer's ligand-bound state [48]. Therefore, the stabilization of the aptamer structure upon nucleotide binding provides an effective switching mechanism for gene control.

A structural model of the related aptamer domain from the adenine-specific riboswitch of the *Vibrio vulnificus add* gene also has been reported [44••]. One of the most striking features of the guanine and adenine riboswitches is that they adopt nearly identical tertiary structures despite sharing less than 60% sequence identity (Figure 3a,b). Likewise, the binding compartment formed by the adenine-specific aptamer is nearly identical to that observed for the *xpt-pbuX* aptamer (Figure 3e). However, the most important difference in nucleotide sequence between the two riboswitch classes occurs at one of the nucleotides that forms this compartment (position 74 of the *B. subtilis xpt-pbuX* RNA). The nucleotide at this position is a cytosine in guanine-binding riboswitches and a uracil in adenine-binding riboswitches (Figure 3c–e). As previously suggested [36•], each aptamer uses this nucleotide to selectively bind its target purine through the formation of standard Watson–Crick base pairing interactions.

Conclusions

Despite the recent discoveries of new RNA genetic elements, it is likely that the current collection of known elements reflects only a small fraction of the contribution that RNA makes to the regulation of modern cells. Furthermore, it seems likely that the diversity of RNA structure and function could have been harnessed by the earliest life forms to construct sensory and regulatory RNAs. Some riboswitches could be direct descendents of ancient metabolite sensors that first emerged in the RNA world [28]. This hypothesis is certainly intriguing, given that all riboswitches identified to date are triggered by compounds that are near universal in their evolutionary distribution, and that many of these ligands carry phosphate and nucleotide-like moieties, as would be expected of metabolites from an RNA world [50,51].

However, riboswitches need not be remnants of an ancient sensory system to have taken their current place in gene regulation systems. RNA could simply be best suited to perform many of the regulatory roles it serves in modern cells. Many riboswitches appear to be widespread in bacteria and therefore they might have been present in the last common ancestor of bacteria. This does not mean that they were present in a purely RNA world organism, but might have emerged after organisms of the RNA world gave way to protein-dominated life forms. Also, lateral transfer and repetitive re-invention of riboswitch classes could have caused riboswitches to be distributed widely despite possible recent emergence. What has been learned about riboswitches over the past several years is consistent with either hypothesis. Furthermore, it seems reasonable to speculate that there might be a mixture of lineages for riboswitches, whereby some are post RNA world representatives whereas others date back to a time before proteins were present.

Given that some riboswitches sense small molecules cooperatively and others self-cleave upon ligand binding, it is possible that future discoveries will expand our understanding of the capabilities of RNA. As for the known riboswitch classes, there remain many questions of structure and mechanism to be addressed. The recent crystal structures have helped in this regard, lending insight into how metabolites are recognized and how the binding domains discriminate between similar molecules. For the guanine and adenine aptamers, recognition is the result of a simple Watson–Crick base pairing interaction. One might speculate that similar strategies are employed, at least in part, for the other nucleotide metabolites, such as FMN and SAM. However, it is less easily predicted how an aptamer will form a binding pocket for a simple molecule such as glycine or lysine. Additional structural and biophysical studies will be needed to address some of the more complex mechanistic questions, such as those concerning the interactions between aptamer and expression platform, between aptamer and aptamer when cooperative function is present, or between the expression platform and RNA polymerase.

One of the major open questions in the field is how and to what degree riboswitches function in eukaryotes. As mentioned earlier, riboswitches have been found in plants and fungi, where they are located near splice site junctions of introns and appear to be regulating mRNA splicing [27•,52•]. However, more detailed characterization is needed to determine the precise role of riboswitches in these processes. Although only one class of metabolite-binding RNA has been identified in eukaryotes [27•], it is possible that improved bioinformatics and genomics approaches will reveal new riboswitch candidates within eukaryotic genomes. Currently, there is no reason to assume that higher organisms cannot exploit the molecular recognition and allosteric properties of RNA, as do their bacterial counterparts.

Acknowledgements

We would like to acknowledge Mary Stahley for assistance with structure images, and Jeffrey Barrick for discussion and comments on the review. Riboswitch research in the Breaker laboratory is funded by grants from the National Institutes of Health, the National Science Foundation and DARPA (Defense Advanced Research Projects Agency).

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Bartel DP: **MicroRNAs: genomics, biogenesis, mechanism, and function.** *Cell* 2004, **116**:281-297.

2. Erdmann VA, Barciszewska MZ, Hochberg A, de Groot N, Barciszewski J: **Regulatory RNAs.** *Cell Mol Life Sci* 2001, **58**:960-977.

3. He L, Hannon GJ: **MicroRNAs: small RNAs with a big role in gene regulation.** *Nat Rev Genet* 2004, **5**:522-531.

4. Erdmann VA, Barciszewska MZ, Szymanski M, Hochberg A, de Groot N, Barciszewski J: **The non-coding RNAs as riboregulators.** *Nucleic Acids Res* 2001, **29**:189-193.

5. Henkin TM, Yanofsky C: **Regulation by transcription attenuation in bacteria: how RNA provides instructions for transcription termination/antitermination decisions.** *Bioessays* 2002, **24**:700-707.

6. Storz G, Opdyke JA, Zhang A: **Controlling mRNA stability and translation with small, noncoding RNAs.** *Curr Opin Microbiol* 2004, **7**:140-144.

7. Grundy FJ, Henkin TM: **The T box and S box transcription termination control systems.** *Front Biosci* 2003, **8**:d20-d31.

8. Chowdhury S, Ragaz C, Kreuger E, Narberhaus F: **Temperature-controlled structural alterations of an RNA thermometer.** *J Biol Chem* 2003, **278**:47915-47921.

9. Morita M, Kanemori M, Yanagi H, Yura T: **Heat-induced synthesis of sigma32 in *Escherichia coli*: structural and functional dissection of rpoH mRNA secondary structure.** *J Bacteriol* 1999, **181**:401-410.

10. Morita MT, Tanaka Y, Kodama TS, Kyogoku Y, Yanagi H, Yura T: **Translational induction of heat shock transcription factor sigma32: evidence for a built-in RNA thermosensor.** *Genes Dev* 1999, **13**:655-665.

11. Kamath-Loeb AS, Gross CA: **Translational regulation of sigma 32 synthesis: requirement for an internal control element.** *J Bacteriol* 1991, **173**:3904-3906.

12. Winkler WC, Breaker RR: **Genetic control by metabolite-binding riboswitches.** *ChemBioChem* 2003, **4**:1024-1032.

13. Vitreschak AG, Rodionov DA, Mironov AA, Gelfand MS: **Riboswitches: the oldest mechanism for the regulation of gene expression?** *Trends Genet* 2004, **20**:44-50.

14. Lai EC: **RNA sensors and riboswitches: self-regulating messages.** *Curr Biol* 2003, **13**:R285-R291.

15. Grundy FJ, Henkin TM: **Regulation of gene expression by effectors that bind to RNA.** *Curr Opin Microbiol* 2004, **7**:126-131.

16. Brantl S: **Bacterial gene regulation: from transcription attenuation to riboswitches and ribozymes.** *Trends Microbiol* 2004, **12**:473-475.

17. Gelfand MS, Mironov AA, Jomantas J, Kozlov YI, Perumov DA: **A conserved RNA structure element involved in the regulation of bacterial riboflavin synthesis genes.** *Trends Genet* 1999, **15**:439-442.

18. Nou X, Kadner RJ: **Adenosylcobalamin inhibits ribosome binding to btuB RNA.** *Proc Natl Acad Sci USA* 2000, **97**:7190-7195.

19. Stormo GD, Ji Y: **Do mRNAs act as direct sensors of small molecules to control their expression?** *Proc Natl Acad Sci USA* 2001, **98**:9465-9467.

20. Mandal M, Boese B, Barrick JE, Winkler WC, Breaker RR: **Riboswitches control fundamental biochemical pathways in *Bacillus subtilis* and other bacteria.** *Cell* 2003, **113**:577-586.

21. Mironov AS, Gusarov I, Rafikov R, Lopez LE, Shatalin K, Kreneva RA, Perumov DA, Nudler E: **Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria.** *Cell* 2002, **111**:747-756.

22. Winkler WC, Cohen-Chalamish S, Breaker RR: **An mRNA structure that controls gene expression by binding FMN.** *Proc Natl Acad Sci USA* 2002, **99**:15908-15913.

23. Sudarsan N, Wickiser JK, Nakamura S, Ebert MS, Breaker RR:
• **An mRNA structure in bacteria that controls gene expression by binding lysine.** *Genes Dev* 2003, **17**:2688-2697.
This report, together with [24•,25•], describes the characteristics of a riboswitch that senses lysine, which represents the first example of a non-nucleotide-based riboswitch ligand. Using a combination of *in vitro* binding and transcription assays, *in vivo* reporters and bioinformatics, these three reports show regulation to be exerted by transcription termination.

24. Grundy FJ, Lehman SC, Henkin TM: **The L box regulon: lysine**
• **sensing by leader RNAs of bacterial lysine biosynthesis genes.** *Proc Natl Acad Sci USA* 2003, **100**:12057-12062.
See annotation to [23•].

25. Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS:
• **Regulation of lysine biosynthesis and transport genes in**

bacteria: yet another RNA riboswitch? *Nucleic Acids Res* 2003, **31**:6748-6757.
See annotation to [23•].

26. Winkler W, Nahvi A, Breaker RR: **Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression.** *Nature* 2002, **419**:952-956.

27. • Sudarsan N, Barrick JE, Breaker RR: **Metabolite-binding RNA domains are present in the genes of eukaryotes.** *RNA* 2003, **9**:644-647.
Two reports [27•,52•] describe evidence of riboswitches in eukaryotes. Bioinformatics efforts show TPP riboswitches near the splice site junctions of introns in various plants and fungi, implying riboswitch control at the level of RNA processing.

28. Soukup JK, Soukup GA: **Riboswitches exert genetic control through metabolite-induced conformational change.** *Curr Opin Struct Biol* 2004, **14**:344-349.

29. Nudler E, Mironov AS: **The riboswitch control of bacterial metabolism.** *Trends Biochem Sci* 2004, **29**:11-17.

30. Mandal M, Breaker RR: **Gene regulation by riboswitches.** *Nat Rev Mol Cell Biol* 2004, **5**:451-463.

31. Lee JF, Hesselberth JR, Meyers LA, Ellington AD: **Aptamer database.** *Nucleic Acids Res* 2004, **32**:D95-D100.

32. Ellington AD, Szostak JW: ***In vitro* selection of RNA molecules that bind specific ligands.** *Nature* 1990, **346**:818-822.

33. Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS: **Comparative genomics of thiamin biosynthesis in prokaryotes. New genes and regulatory mechanisms.** *J Biol Chem* 2002, **277**:48949-48959.

34. Gusarov I, Nudler E: **The mechanism of intrinsic transcription termination.** *Mol Cell* 1999, **3**:495-504.

35. Nudler E, Gusarov I: **Analysis of the intrinsic transcription termination mechanism and its control.** *Methods Enzymol* 2003, **371**:369-382.

36. • Mandal M, Breaker RR: **Adenine riboswitches and gene activation by disruption of a transcription terminator.** *Nat Struct Mol Biol* 2004, **11**:29-35.
The first characterization of a riboswitch that activates, rather than inhibits, gene expression in response to metabolite. The riboswitch binds adenine and is structurally related to a previously characterized guanine-binding riboswitch [20].

37. Johansen LE, Nygaard P, Lassen C, Agerso Y, Saxild HH: **Definition of a second *Bacillus subtilis* pur regulon comprising the pur and xpt-pbuX operons plus pbuG, nupG (yxjA), and pbuE (ydhL).** *J Bacteriol* 2003, **185**:5200-5209.

38. Barrick JE, Corbino KA, Winkler WC, Nahvi A, Mandal M, Collins J, Lee M, Roth A, Sudarsan N, Jona I *et al.*: **New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control.** *Proc Natl Acad Sci USA* 2004, **101**:6421-6426.

39. • Winkler WC, Nahvi A, Roth A, Collins JA, Breaker RR: **Control of gene expression by a natural metabolite-responsive ribozyme.** *Nature* 2004, **428**:281-286.
Characterization of a novel metabolite-responsive ribozyme. Cleavage activity of the *glmS* element is shown to be activated upon binding to GlcN6P. Furthermore, *in vitro* cleavage activity is shown to correlate with *in vivo* regulation of a reporter gene.

40. •• Mandal M, Lee M, Barrick JE, Weinberg Z, Emilsson GM, Ruzzo WL, Breaker RR: **A glycine-dependent riboswitch that uses cooperative binding to control gene expression.** *Science* 2004, **306**:275-279.
A riboswitch responsive to glycine is reported. Surprisingly, the riboswitch is found to utilize two aptamer domains to bind glycine cooperatively, a trait normally associated with proteins.

41. Hill AV: **The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves.** *J Physiol* 1910, **40**:iv-vii.

42. Edelstein SJ: **Cooperative interactions of hemoglobin.** *Annu Rev Biochem* 1975, **44**:209-232.

43. •• Batey RT, Gilbert SD, Montange RK: **Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine.** *Nature* 2004, **432**:411-415.
The first reported high-resolution structure of a riboswitch aptamer domain. The guanine riboswitch aptamer is co-crystallized with hypoxanthine, a metabolite of the guanine biosynthetic pathway previously shown to regulate the riboswitch.

44. •• Serganov A, Yuan YR, Pikovskaya O, Polonskaia A, Malinina L, Phan AT, Hobartner C, Micura R, Breaker RR, Patel DJ: **Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs.** *Chem Biol* 2004, **11**:1729-1741.
This report includes high-resolution structures of both the guanine and adenine riboswitch aptamers bound to guanine and adenine, respectively. The structures reveal how discrimination between the purine metabolites is achieved, despite the two aptamers having near-identical tertiary structure.

45. Christiansen LC, Schou S, Nygaard P, Saxild HH: **Xanthine metabolism in *Bacillus subtilis*: characterization of the xpt-pbuX operon and evidence for purine- and nitrogen-controlled expression of genes involved in xanthine salvage and catabolism.** *J Bacteriol* 1997, **179**:2540-2550.

46. Flinders J, DeFina SC, Brackett DM, Baugh C, Wilson C, Dieckmann T: **Recognition of planar and nonplanar ligands in the malachite green-RNA aptamer complex.** *ChemBioChem* 2004, **5**:62-72.

47. Hermann T, Patel DJ: **Adaptive recognition by nucleic acid aptamers.** *Science* 2000, **287**:820-825.

48. Lescoute A, Westhof E: **Riboswitch structures: purine ligands replace tertiary contacts.** *Chem Biol* 2005, **12**:10-13.

49. Baugh C, Grate D, Wilson C: **2.8 Å crystal structure of the malachite green aptamer.** *J Mol Biol* 2000, **301**:117-128.

50. Jeffares DC, Poole AM, Penny D: **Relics from the RNA world.** *J Mol Evol* 1998, **46**:18-36.

51. White HB III: **Coenzymes as fossils of an earlier metabolic state.** *J Mol Evol* 1976, **7**:101-104.

52. • Kubodera T, Watanabe M, Yoshiuchi K, Yamashita N, Nishimura A, Nakai S, Gomi K, Hanamoto H: **Thiamine-regulated gene expression of *Aspergillus oryzae* thiA requires splicing of the intron containing a riboswitch-like domain in the 5′-UTR.** *FEBS Lett* 2003, **555**:516-520.
This report describes testing for function one of the eukaryotic riboswitches from *Aspergillus oryzae* and shows that control might be at the level of splicing.

Published online 9 July 2007 *Nucleic Acids Research, 2007, Vol. 35, No. 14* **4809–4819**
doi:10.1093/nar/gkm487

Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline

Zasha Weinberg[1,*], Jeffrey E. Barrick[2,3], Zizhen Yao[4], Adam Roth[2], Jane N. Kim[1], Jeremy Gore[1], Joy Xin Wang[1,2], Elaine R. Lee[1], Kirsten F. Block[1], Narasimhan Sudarsan[1], Shane Neph[5], Martin Tompa[4,5], Walter L. Ruzzo[4,5] and Ronald R. Breaker[1,2,3]

[1]Department of Molecular, Cellular and Developmental Biology, [2]Howard Hughes Medical Institute, [3]Department of Molecular Biophysics and Biochemistry, Yale University, Box 208103, New Haven, CT 06520-8103, USA [4]Department of Computer Science and Engineering and [5]Department of Genome Sciences, University of Washington, Box 352350, Seattle, WA 98195-2350, USA

Received April 25, 2007; Revised June 4, 2007; Accepted June 5, 2007

ABSTRACT

We applied a computational pipeline based on comparative genomics to bacteria, and identified 22 novel candidate RNA motifs. We predicted six to be riboswitches, which are mRNA elements that regulate gene expression on binding a specific metabolite. In separate studies, we confirmed that two of these are novel riboswitches. Three other riboswitch candidates are upstream of either a putative transporter gene in the order Lactobacillales, citric acid cycle genes in Burkholderiales or molybdenum cofactor biosynthesis genes in several phyla. The remaining riboswitch candidate, the widespread Genes for the Environment, for Membranes and for Motility (GEMM) motif, is associated with genes important for natural competence in *Vibrio cholerae* and the use of metal ions as electron acceptors in *Geobacter sulfurreducens*. Among the other motifs, one has a genetic distribution similar to a previously published candidate riboswitch, *ykkC/yxkD*, but has a different structure. We identified possible non-coding RNAs in five phyla, and several additional *cis*-regulatory RNAs, including one in ε-proteobacteria (upstream of *purD*, involved in purine biosynthesis), and one in Cyanobacteria (within an ATP synthase operon). These candidate RNAs add to the growing list of RNA motifs involved in multiple cellular processes, and suggest that many additional RNAs remain to be discovered.

INTRODUCTION

Recent discoveries of novel structured RNAs (1–4) indicate that such RNAs are common in cells. To assist in discovering additional structured RNAs, we have developed an automated pipeline that can identify conserved RNAs within bacteria (5). This pipeline assembles the potential 5′ untranslated regions (UTRs) of mRNAs of homologous genes, and uses the CMfinder (6) program to predict conserved RNA structures, or 'motifs', within each set of UTRs. Automated homology searches are then employed to find additional examples of these motifs, which CMfinder uses to improve the secondary structure model and sequence alignment of each motif. The output is a set of alignments with a predicted RNA secondary structure, and these alignments are subsequently analyzed manually to make improvements to the model and to assess which motifs merit further study.

Although other automated searches for RNAs have been performed (7–15), our pipeline (5) is distinguished from these by three features. First, our pipeline uses CMfinder, which can discover a motif even when some input sequences either do not contain the motif or include motif representatives carrying unrelated sequence domains. CMfinder also can produce a useful alignment even with low sequence conservation, however, the algorithm will exploit whatever sequence similarity is present. Second, our pipeline integrates homology searches to automatically refine the alignment and structural model for each motif. Third, since our pipeline aligns UTRs of homologous genes, it is well suited to find *cis*-regulatory RNAs, and

*To whom correspondence should be addressed. Tel: (203) 432-6554; Fax: (203) 432-6161; Email: zasha.weinberg@yale.edu
Correspondence may also be addressed to Ronald R. Breaker. Tel: (203) 432-9389; Fax: (203) 432-6161; Email: ronald.breaker@yale.edu
Present address:
Jeffrey E. Barrick, Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI 48824, USA

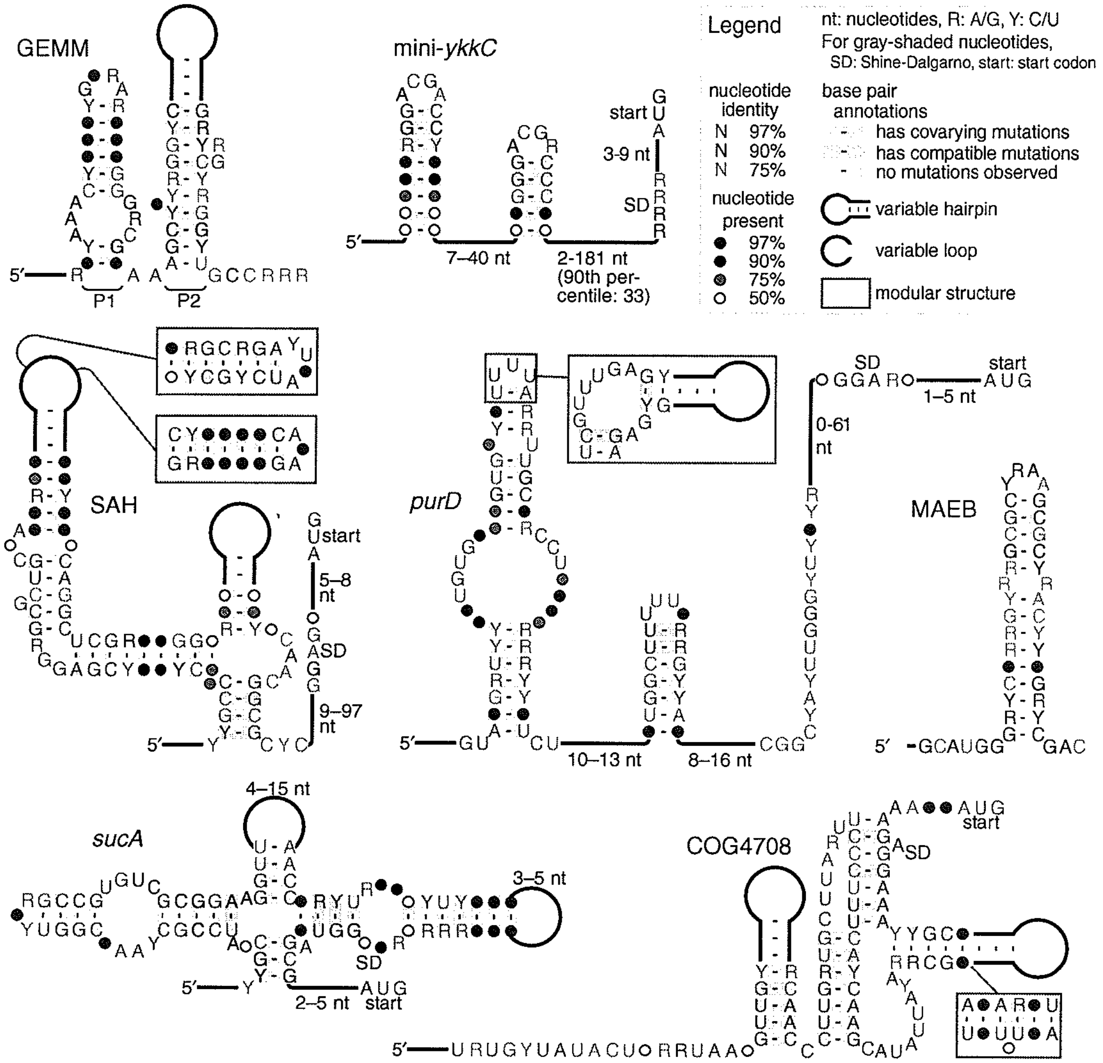


Figure 1. Consensus sequences and structures are depicted for seven of the 22 motifs identified. Other motifs are presented as Supplementary Data, as are the alignments on which these diagrams are based. Calculations for conservation of nucleotide identity/presence and evidence of covariation are described in the 'Materials and methods' section. Proposed base pairs with more than 5% non-canonical or missing nucleotides are not classified as covarying. Note that the levels of nucleotide conservation are affected both by biochemical constraints on the motif and by phylogenetic diversity; motifs with limited range (e.g. the COG4708 motif) will appear more conserved. Some covarying positions in variable-length stems are not shown.

its dependence on sequence conservation is further reduced.

Indeed, using the bacterial phylum Firmicutes as a test case, we previously demonstrated that this pipeline makes useful predictions for virtually all known *cis*-regulatory RNAs (5). The pipeline also finds motifs that are likely to be *trans*-encoded, or 'non-coding RNAs' (ncRNAs), when these happen to be upstream of homologous genes.

In the present report, we describe the use of this pipeline to find structured RNAs amongst all bacteria whose genomes have been sequenced. We describe 22 novel motifs that are likely to be conserved, structured RNAs. We were particularly interested in discovering riboswitches, a type of structured RNA usually found in mRNAs that directly senses a specific small molecule and regulates gene expression (16,17). Subsequent experiments have confirmed that two of these motifs are novel riboswitches. The first binds SAH (*S*-adenosylhomocysteine) (J.X.W., D. Rivera, E.R.L., R.R.B., in preparation; Figure 1) and the second is a SAM (*S*-adenosylmethionine)-binding riboswitch in *Streptomyces coelicolor* and related species (Z.W., E.E. Regulski, R.R.B., unpublished data). Experimental evidence with another riboswitch candidate indicates that it senses molybdenum cofactor or 'Moco' (E.E. Regulski, R. Moy, R.R.B., unpublished data).

A candidate of particular interest is the Genes for the Environment, for Membranes, and for Motility (GEMM) motif, which has properties that are typical of riboswitches. For example, GEMM is a widespread and highly

Table 1. Summary of putative structured RNA motifs

Motif	RNA?	Cis?	Switch?	Phylum/class	M,V	Cov.	#	Non-cis
GEMM	Y	Y	y	Widespread	V	21	322	12/309
Moco	Y	Y	Y	Widespread	M,V	15	105	3/81
SAH	Y	Y	Y	Proteobacteria	M,V	22	42	0/41
SAM-IV	Y	Y	Y	Actinobacteria	V	28	54	2/54
COG4708	Y	Y	y	Firmicutes	M,V	8	23	0/23
sucA	Y	Y	y	β-proteobacteria		9	40	0/40
23S-methyl	Y	y	n	Firmicutes		12	38	1/37
hemB	Y	?	?	β-proteobacteria	V	12	50	2/50
(anti-*hemB*)		(n)	(n)				(37)	(31/37)
MAEB	?	Y	n	β-proteobacteria		3	662	15/646
mini-*ykkC*	Y	Y	?	Widespread	V	17	208	1/205
purD	y	y	?	ε-proteobacteria	M	16	21	0/20
6C	y	?	n	Actinobacteria		21	27	1/27
alpha-transposases	?	N	N	α-proteobacteria		16	102	39/99
excisionase	?	?	n	Actinobacteria		7	27	0/27
ATPC	y	?	?	Cyanobacteria		11	29	0/23
Cyano-30S	Y	Y	n	Cyanobacteria		7	26	0/23
lacto-1	?	?	n	Firmicutes		10	97	18/95
lacto-2	y	N	n	Firmicutes		14	357	67/355
TD-1	y	?	n	Spirochaetes	M,V	25	29	2/29
TD-2	y	N	n	Spirochaetes	V	11	36	17/36
coccus-1	?	N	N	Firmicutes		6	246	112/189
gamma-150	?	N	N	γ-proteobacteria		9	27	6/27

'RNA' = functions as RNA (as opposed to dsDNA), 'Cis' = *cis*-regulatory, 'Switch' = riboswitch. Evaluation: 'Y' = certainly true, 'y' = probably true, '?' = possible, 'n' = probably not, 'N' = certainly not. These evaluations were conducted prior to experimental examinations. Criteria for classification as an RNA include evidence of covariation and variable-length or modular stems. Evidence of covariation is strongest with covarying nucleotide positions for which surrounding sequence conservation permits high confidence that the covarying positions are correctly aligned, and was assessed manually based on alignments. Probable *cis*-regulatory motifs were consistently located upstream of homologous genes, or a set of genes with related functions, and often had features typical of known gene-control mechanisms (a transcription terminator or stem sequestering the Shine–Dalgarno sequence). Likely riboswitches were motifs that were classified as an RNA and a *cis*-regulatory element, showed evidence of high conservation of nucleotides at some positions, exhibited a complex secondary structure (not just a hairpin) and were associated with genes that were judged likely to be controlled by a small molecule. Motifs are characterized in detail according to these criteria in Supplementary Data. Remaining columns are 'Phylum/class' (phylum containing the motif, or class for Proteobacteria), 'M,V' ('M' = has modular stems, which are stems that are only sometimes present, 'V' = variable-length stems), 'Cov.' = number of covarying paired positions (see 'Methods' section; note that it is not advisable to rank motifs solely by this number, but rather the alignment as a whole should be evaluated), '#' = number of representatives, 'Non-cis' = X/Y where X is number of representatives that are *not* in a 5′ regulatory configuration to a gene and Y is the number of representatives within sequences that have annotated genes (some RefSeq sequences lack annotations). Moco and SAM-IV riboswitch data will be presented in future reports. Gamma-150 and coccus-1 are only in the supplement.

conserved genetic element that, in 297 out of 309 cases, is positioned such that it is likely to be present in the 5′ UTR of the adjacent open reading frame (ORF). The genes putatively regulated by GEMM are typically related to sensing and reacting to extracellular conditions, which suggests that GEMM might sense a metabolite produced for signal transduction or for cell–cell communication.

Some characteristics of all 22 predicted RNA motifs are summarized in Table 1, and we discuss the features and possible biological roles of some candidates in the 'Results' section. Additional information on all candidates, including annotated multiple sequence alignments, and collections of taxonomy and nearby genes, are presented in Supplementary Data. Raw pipeline predictions are accessible at http://bliss.biology.yale.edu/cmfinder_pipeline_output.

MATERIALS AND METHODS

Identification of candidate RNA motifs

The potential 5′ UTRs of genes classified in the Conserved Domain Database (18) version 2.08 were used as input for our computational pipeline (5). Completed genome sequences and gene positions were taken from RefSeq (19) version 14, but we eliminated genomes whose gene content was highly similar to other genomes. Our UTR extraction algorithm (20) accounted for the fact that a UTR might not always be immediately upstream of the gene due to operon structure.

For each conserved domain, the collected sequences of potential UTRs were given as input to CMfinder (6) version 0.2, which produced local multiple sequence alignments of structurally conserved motifs within the UTR sets. These alignments were then used to search for additional homologs within annotated intergenic regions, except that intergenic regions were extended by 50 nt on both ends to account for misannotated ORFs. This homology search was performed with the RaveNnA program version 0.2f (21–23) with ML-heuristic filters (23), Covariance Model (24) in global mode implemented by Infernal (25) version 0.7 and an E-value (26) cutoff of 10. CMfinder then refined its initial alignment using these new homologs. Known RNAs were detected based on the Rfam Database (27) version 7.0. Predictions were scored based on phylogenetic conservation of short sequences (20), diversity of species and structural criteria. The pipeline algorithm is described in more detail elsewhere (5). The Supplementary Data includes a list of software and databases used.

We split bacterial genomic sequence data into groups, and performed UTR extraction, motif prediction and homology search on each group separately. This was motivated by the fact that UTRs in different phyla often do not contain the same motif. However, phyla with few sequenced members were coalesced based on taxonomy to try to assemble sufficient sequence data to predict a motif. Phyla with only one or two sequenced members (e.g. Chloroflexi) were ignored. The following eight groups were used: (1) Firmicutes, (2) Actinobacteria, (3) α-proteobacteria, (4) β-proteobacteria, (5) γ-proteobacteria, (6) δ- and ε-proteobacteria, (7) Cyanobacteria and (8) Bacteroidetes, Chlorobi, Chlamydiae, Verrucomicrobia and Spirochaetes.

Analysis of candidate motifs and homology search strategies

We then selected promising motifs, further analyzed them by performing additional homology searches and edited their alignments with RALEE (28). We used NCBI BLAST (29), Mfold (30), RnaII (31) (to identify rho-independent transcription terminators), CMfinder and RaveNnA to assist in these analyses. Information on metabolic pathways associated with motifs was retrieved from KEGG (32). To expand alignments by identifying additional structured elements, we extended alignments on their 5′ and 3′ ends by 50–100 nt, and realigned as necessary using either CMfinder or by manual inspection.

The additional searches for homologs used RaveNnA in several ways. We found both global- and local-mode (25) Covariance Model searches gave complementary results. Sequence databases used in manually directed searches were the ‘microbial’ subset of RefSeq version 19 (19), and environmental shotgun sequences from acid mine drainage (33) (GenBank accession AADL01000000) and Sargasso Sea (34) (AACY01000000). Additional marine shotgun sequences were used for the *sucA* and ATPC motifs (35).

Because the full set of sequences is roughly 3.2 billion nucleotides, searches can report many false positives, especially for shorter motifs. When appropriate, we searched four kinds of subsets of these sequences. First, we did not always use the environmental sequence data. Second, we sometimes searched only genomes in the bacterial group (e.g. β-proteobacteria) from which the motif was originally derived. Third, we sometimes searched only intergenic regions (extended by 50 nt as before). Fourth, we used the BLAST program tblastn to search for genes homologous to those associated with the motif. RaveNnA searches were then conducted on the 2 kb upstream of these matches (which is expected to contain the 5′ UTR), and 200 nt downstream (since apparent coding homology might extend upstream of the true ORF, causing BLAST to misidentify the start codon). Using the motif's bacterial group as the BLAST database facilitated the discovery of highly diverged homologs. For example, a search upstream of *purD* genes in ε-proteobacteria revealed homologs of the *purD* motif (see later) with a truncated stem. Additionally, with such small databases, we can forego the ML-heuristic filters in RaveNnA. When the full sequence set is used as the BLAST database, it can help to find homologs in other phyla.

Rejection of motifs

Motifs were rejected from further study when they failed to show features that are characteristic of structured RNAs. To help reject motifs with spurious predictions of structure, we performed homology searches using sequence information only, by removing all base pairs in the predicted structure. Sequence-based matches that do not conserve the structure indicate that the predicted structure is incorrect. However, such homologs can be missed by Covariance Models, which assume the structure is conserved. Several motifs were rejected using this strategy when sequence homologs revealed that the proposed structure was, in fact, poorly conserved.

We also generally rejected repetitive elements, which we defined as elements appearing many times per genome and showing extremely high sequence conservation, but little structure conservation. Although some of these repetitive elements could correspond to structured RNAs, there is a little support for such a hypothesis without good evidence of covariation.

Establishing the extent of conservation and covariation for consensus diagrams

To establish the extent of conservation reflected in consensus diagrams (e.g. in Figure 1), sequences were weighted to de-emphasize highly similar homologs. Weighting used the GSC algorithm (36), as implemented by Infernal (25), and weighted nucleotide frequencies were then calculated at each position in the multiple sequence alignment. To classify base pairs as covarying, the weighted frequency of Watson–Crick or G–U pairs was calculated. However, aligned sequences in which both nucleotides were missing or where the identity of either nucleotide was uncertain (e.g. was ‘N’, signifying any of the four bases) were discarded. Classification as a covarying position was made if two sequences had Watson–Crick or G–U pairs that differ at both positions amongst sequences that carry the motif. If only one position differed, the occurrence was classified as a compatible mutation. However, if the frequency of non-Watson–Crick or G–U pairs was more than 5%, we did not annotate these positions as covarying or as compatible mutations.

RESULTS

Evaluation and analysis of novel RNA motifs

Promising structured RNA motifs predicted by the CMfinder pipeline were examined manually to refine the consensus sequence and structural models (see ‘Materials and methods’ section) and to provide information on possible function. Key findings for each candidate are summarized in Table 1. Of the 22 motifs identified, seven are depicted in Figure 1 and the remainder are depicted in Supplementary Data.

Our decision to select a candidate RNA motif for further study was based on a qualitative evaluation of conservation of both sequence and structure, covariation and gene context (e.g. whether or not the motif is consistently upstream of a specific gene family). Conserved sequence is important because structured RNAs usually have many conserved nucleotides in regions that form complex tertiary structures or are under other constraints. Structured RNAs sometimes have variable-length stems or 'modular stems' (stems that are present in some but not all representatives). The fact that both sides of a stem either appear together or neither appear is analogous to covariation, and is evidence that the structure is conserved.

Cis-regulatory RNAs such as riboswitches are regions of mRNAs that regulate gene expression. In bacteria, most *cis*-regulatory RNAs occur in the 5′ UTR of the mRNA under regulation. Although it is not possible to reliably predict the transcription start site, we declare representatives of a motif as positioned in a '5′ regulatory configuration' to a gene when the element could be in the 5′ UTR of an mRNA (if the transcription start site is 5′ to the element). When most or all representatives of a motif are in a 5′ regulatory configuration to a gene, this is evidence that the motif might have a *cis*-regulatory function.

Cis-regulatory RNAs often have one of two noteworthy structural features: rho-independent transcription terminators, or stems that overlap the Shine–Dalgarno sequence (bacterial ribosome-binding site) (16). Rho-independent transcription terminators usually consist of a strong hairpin followed by four or more U residues (37). Regulatory RNA domains can control gene expression by conditionally forming the terminator stem. Similarly, conditionally formed stems can overlap the Shine–Dalgarno sequence, thereby regulating genes at the translational level (16).

Some motifs identified in this study consist of a single or tandem hairpins. It is possible that some of these are protein-binding motifs in which a homodimeric protein binds to a given DNA-based element in opposite strands. For convenience, we also describe such motifs as having a 5′ regulatory configuration, even though they might not form structured RNAs or function at the mRNA level.

The GEMM motif

GEMM is widespread in bacteria and appears to have a highly conserved sequence and structure suggestive of a function that imposes substantial biochemical constraints on the putative RNA. We found 322 GEMM sequences in both Gram-positive and Gram-negative bacteria. It is common in δ-proteobacteria, particularly in *Geobacter* and related genera. Within γ-proteobacteria, it is ubiquitous in Alteromonadales and Vibrionales. It is also common in certain orders of the phyla Firmicutes and Plantomycetes. Prominent pathogens with GEMM include the causative agents of cholera and anthrax.

Out of 309 GEMM instances where sequence data includes gene annotations, GEMM is in a 5′ regulatory configuration to a gene in 297 cases, implying a *cis*-regulatory role. Genes presumably regulated by GEMM display a wide range of functions, but most genes relate to the extracellular environment or to the membrane, and many are related to motility.

GEMM consists of two adjacent hairpins (paired regions) designated P1 and P2 (Figure 1). P1 is highly conserved in sequence and structure, and consists of 2- and 6-bp stems separated by a 3-nt internal loop and capped by a terminal loop. The internal loop is highly conserved, and the terminal loop is almost always a GNRA tetraloop (38). The P1 stem exhibits considerable evidence of covariation at several positions, and is highly conserved in structure over a wide range of bacteria. This fact, and the more modest covariation and variable-length stems of P2, provide strong evidence that GEMM functions as a structured RNA. The sequence linking P1 and P2 is virtually always AAA, with only two exceptions in 322 examples.

The P2 hairpin shows more modest conservation than P1. When the P1 tetraloop is GAAA, a GNRA tetraloop receptor usually appears in P2. This receptor is often the well-known 11-nt motif, which might be favored by GAAA loops (39), but some sequences could be novel tetraloop receptors. When P1 has a GYRA tetraloop, the receptor-like sequence is almost never present, although a bulge nearer the P2 base is sometimes found (Figure 2).

Many instances of GEMM include a rho-independent transcription terminator hairpin. The 5′ side of the terminator stem often overlaps (and presumably competes with) the 3′ side of the P2 stem (Figure 2B). If GEMM is a riboswitch, ligand binding could stabilize the proposed P1 and P2 structure, thus preventing the competing transcription terminator from forming. In this model, higher ligand concentrations will increase gene expression. One third of GEMM representatives in δ-proteobacteria, and some in other taxa, are in a 'tandem' arrangement, wherein one instance appears 3′ and nearby to another in the same UTR. Such arrangements of regulatory RNAs are implicated in more sophisticated control of gene expression than is permitted by a single regulatory RNA configuration (40–42).

An understanding of the biological role of GEMM will likely shed light on the broad variety of microbial processes that it appears to regulate. In fact, GEMM is implicated in two systems that are already the object of several studies in the species *Vibrio cholerae* and in *Geobacter sulfurreducens*. *Vibrio cholerae* causes cholera in humans, but spends much of its lifecycle in water, where it can adhere to chitin-containing exoskeletons of many crustaceans. Chitin, a polymer of GlcNAc (*N*-acetylglucosamine), has been shown to affect expression of many *V. cholerae* genes (43). GEMM appears to regulate two of these chitin-induced genes. The first, *gbpA*, is important for adhering to chitin beads (43) and human epithelial cells (44), as well as infection of mice (44).

The second chitin-induced gene is $tfoX^{VC}$. Remarkably, chitin induces natural competence in *V. cholerae* (45), and $tfoX^{VC}$ expression is essential for this competence.

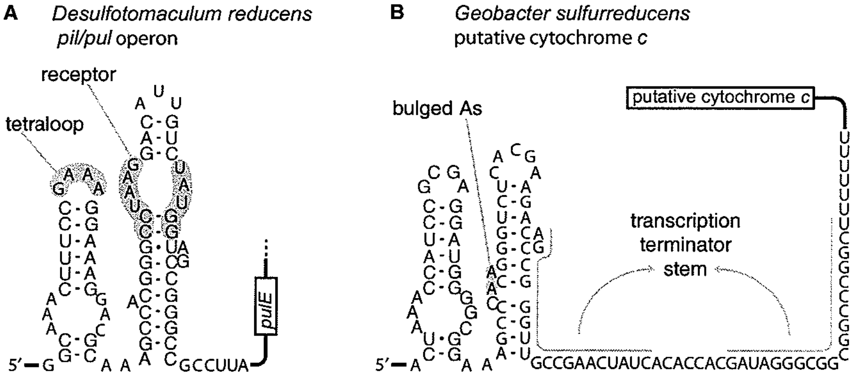


Figure 2. Common features of GEMM motifs. Two GEMM instances were selected to illustrate common features, although these two examples do not represent the full 322 GEMMs (see Supplementary Data). (A) This putative RNA contains a canonical GNRA tetraloop and receptor (gray regions). Almost 50% of GEMM instances contain a likely tetraloop receptor. Only the first gene in the downstream operon is shown. (B) Some GEMM RNAs lack the tetraloop receptor, but there are two extra bulged A residues (gray shading) that are found in roughly half of the sequences lacking a receptor. Gray overlined nucleotides can fold to form a stem of a rho-independent transcription terminator (followed by 3′-trailing Us). This terminator appears to compete with the 3′ part of the P2 stem (right-most hairpin). 78 of 322 GEMM instances have predicted transcription terminators overlapping P2.

Table 2. Gene families that appear to be regulated by GEMM in more than one instance

Functional role	Gene families
Pili and flagella	*cpaA*, *flgB*, *flgC*, *flgG*, *fliE*, *fliG*, *fliM*, *motA*, *motB*, *papC*, *papD*, *pilM*, *pilO*, *pilQ*, *pulF*
Secretion (related to pili/flagella)	*fhaC*, *fliF*, *fliI*, *hofQ*
Chemotaxis regulator	*cheW*, *cheY*, methyl-accepting chemotaxis protein, Cache domain (classically associated with chemotaxis receptors in bacteria)
Signal transduction	PAS domain, histidine kinase, HAMP (Histidine kinases, Adenylyl cyclases, Methyl binding proteins, Phosphatases), HD-GYP domain, GGDEF domain
Chitin	chitin/cellulose binding domain, chitinase, carbohydrate-binding protein
Membranes	lysin domain (involved in cell wall remodeling, but might have general peptidoglycan binding function), uncharacterized outer membrane proteins and lipoproteins, putative collagen binding protein
Peptides	non-ribosomal peptide synthase, condensation domain (synthesis of peptide antibiotics), transglutaminase-like cysteine protease, subtilase (superfamily of extracellular peptidases)
Other	*tfoX* (regulator of competence), cytochrome *c*

V. cholerae has two genes that match the CDD models COG3070 and pfam04994 that correspond to separate *tfoX* domains. Both domains yield RPSBLAST (46) E-values better than 10^{-25}. One of these is $tfoX^{VC}$ (locus VC1153). GEMM appears to regulate the other, which we call $tfoX^{GEMM}$ (VC1722). Thus, in *V. cholerae* and related bacteria, GEMM might participate in chitin-induced competence, or even regulate competence in environments not containing elevated chitin concentrations.

Geobacteria sulfurreducens and related δ-proteobacteria can generate ATP by oxidizing organic compounds, using metal ions such as Fe(III) as electron acceptors (47). GEMM is associated with pili assembly genes in *Geobacter* species. Pili in *G. sulfurreducens* have been shown to conduct electricity (48), and are thus a part of the process of reducing metal ions.

Moreover, GEMM appears to regulate seven cytochrome *c* genes in *G. sulfurreducens*. Although this bacteria has 111 putative cytochrome *c* genes, five of the seven GEMM-associated genes have been identified in previous studies, and might have special roles. OmcS (Outer-Membrane Cytochrome S) is one of two proteins that are highly abundant on the outer membrane of *G. sulfurreducens*, and is required for reducing insoluble Fe(III) oxide, but not for soluble Fe(III) citrate (49). OmcG and OmcH are necessary for production of OmcB, an essential cytochrome *c* in many conditions (50). OmcA and OmcT are associated with OmcG, OmcH or OmcS. Only four other Omc annotations remain in *G. sulfurreducens* that have no direct GEMM association: OmcB, OmcC, OmcE and OmcF.

Unlike known riboswitches, GEMM is associated with a great diversity of gene functions (Table 2). This observation indicates that, if GEMM is a riboswitch, it is not serving as a typical feedback sensor for control of a metabolic pathway. Rather, GEMM more likely senses a second-messenger molecular involved in signal transduction or possibly cell–cell communication (51). In this model, different bacteria use GEMM and its signaling molecule to control different processes. The fact that many GEMM-associated genes encode signal transduction domains could suggest a mechanism by which many of the signal transduction proteins are regulated. Preliminary biochemical results indicate that GEMM RNAs indeed serve as aptamer components of a new-found riboswitch class (N.S., E.R.L, R.R.B., unpublished data).

The SAH motif

The SAH motif is highly conserved in sequence and structure (Figure 1), showing covariation within predicted stem regions, including modular and variable-length stems. The SAH motif is found in a 5′ regulatory configuration to genes related to SAH (*S*-adenosylhomocysteine) metabolism, primarily in β- and some γ-proteobacteria, and especially the genus *Pseudomonas*. SAH is a part of the *S*-adenosylmethionine (SAM) metabolic cycle, whose main components include the amino acid methionine. SAH is a byproduct of enzymes that use SAM as a cofactor for methylation reactions. Typically, SAH is hydrolyzed into homocysteine and adenosine. Homocysteine is then used to synthesize methionine, and ultimately SAM.

High levels of SAH are toxic to cells because SAH inhibits many SAM-dependent methyltransferases (52). Therefore cells likely need to sense rising SAH concentrations and dispose of this compound before it reaches toxic levels. The genes that the SAH motif associates with are *S*-adenosylhomocysteine hydrolase (*ahcY*), cobalamin-dependent methionine synthase (*metH*) and methylene-tetrahydrofolate reductase (*metF*), which synthesizes a methyl donor used in methionine synthesis. This genetic arrangement of the SAH motif and its high degree of conservation are consistent with a role in sensing SAH and activating the expression of genes whose products are required for SAH destruction. Indeed, biochemical and genetic evidence supports the hypothesis that this motif is an SAH-sensing riboswitch (J.X.W., D. Rivera, E.R.L. and R.R.B., in preparation).

The COG4708 motif

This motif is found upstream of COG4708 genes in some species of *Streptococcus* and in *Lactococcus lactis*, although some instances of the COG4708 gene family in *Streptococcus* lack the putative RNA motif. COG4708 genes are predicted to encode membrane proteins.

Although the COG4708 motif is highly constrained phylogenetically and has only six unique sequences, it shows covariation, modular stems and variable-length stems (Figure 1). The motif has a pseudoknot that overlaps the putative Shine–Dalgarno sequences of COG4708 genes, which suggests that the motif encodes a *cis*-regulator of these genes.

We recently characterized a riboswitch that senses the modified nucleobase preQ$_1$ (53). Since this riboswitch is associated with COG4708, we proposed that COG4708 is a transporter of a metabolite related to preQ$_1$. Therefore, we hypothesize that the COG4708 motif is also a preQ$_1$-sensing riboswitch. Preliminary experiments support this hypothesis (M. Meyer, A.R. and R.R.B., unpublished data). The COG4708 motif shares no similarity in sequence or structure with the previously characterized preQ$_1$-sensing riboswitch (53).

The *sucA* motif

The *sucA* motif is only found in a 5′ regulatory configuration to *sucA* genes, which are likely co-transcribed with the related downstream genes *sucB*/*aceF* and *lpd*. The products of these three genes synthesize succinyl-CoA from 2-oxoglutarate in the citric acid cycle. All detected instances of the *sucA* motifs are in β-proteobacteria in the order Burkholderiales. Although many nucleotides in the *sucA* motif are strictly conserved, those that are not show covariation and contain very few non-canonical base pairs (Figure 1). The motif has stems that overlap the putative Shine–Dalgarno sequence, so the *sucA* motif probably corresponds to a *cis*-regulatory RNA. Note that the exact position of the putative Shine–Dalgarno sequence is inconsistent among *sucA* motif instances, so is not well reflected in Figure 1 (see alignment in Supplementary Data). The relatively complex structure of the *sucA* motif suggests that it might be a riboswitch. However, it is difficult to evaluate its degree of sequence and structure conservation since the motif is not broadly distributed.

The 23S-methyl motif

This motif is consistently upstream of genes annotated as rRNA methyltransferases that probably act on 23S rRNA. The one exception occurs when 23S-methyl RNA is roughly 3 kb from the 23S rRNA methyltransferase ORF, with other genes on the opposite strand in the intervening sequence. The 23S-methyl motif is confined to Lactobacillales, which is an order of Firmicutes.

The 23S-methyl motif consists of two large hairpins. The second hairpin ends in a run of Us and appears to be a rho-independent transcription terminator. Both stems have considerable covariation, providing strong evidence that they are part of a functional RNA. Although the structural model shows that many paired positions sometimes have non-canonical base pairs, each instance of the motif consists predominantly of energetically favorable pairs, as shown in Supplementary Data. The presence of a putative transcription terminator suggests that this is a *cis*-regulatory RNA. Since 23S rRNA methyltransferase interacts with an RNA substrate, it might autoregulate its expression using the 23S-methyl motif, in a manner similar to autoregulation of ribosomal protein genes (54).

The *hemB*/anti-*hemB* motif

This motif is found in a variety of β-proteobacteria, especially *Burkholderia*. There is some ambiguity as to the DNA strand from which it might be transcribed, because its structure exhibits comparable covariation and conservation in both directions. In one direction, it is often upstream of *hemB* genes. It could be a *cis*-regulatory RNA in this direction, but there are two genes, not homologous to each other, that are immediately downstream of *hemB* motif representatives and these are positioned on the wrong strand to be controlled in the usual manner of *cis*-regulatory RNAs. In the other direction (anti-*hemB*), the motif is not typically in the 5′ UTRs of genes. The anti-*hemB* motif ends in a transcription terminator hairpin. Many genes downstream of anti-*hemB* instances are on the opposite strand, therefore we propose that anti-*hemB* could encode a non-coding RNA.

MAEB (metabolism-associated element in *Burkholderia*)

This motif consists of a single hairpin with several conserved positions (Figure 1). It is widespread in *Burkholderia*, a genus of β-proteobacteria. It typically occurs multiple times in succession (2–6 copies) with conserved linker sequences, but ranges to as many as 12 copies in two instances. In 141 occurrences of single or repetitive MAEB motifs, 132 are in a 5′ regulatory configuration to a gene. In fact, many of these genes are directly involved in primary metabolism (e.g. genes involved in biosynthesis, catabolism or transport of small molecules), and not genes such as DNA repair, replication, signaling or motility. Out of the 46 conserved domains (excluding hypothetical genes) downstream of MAEB in more than one instance, at least 42 are annotated as participating in primary metabolism (see Supplementary Data for list). There are many bacterial genes not involved in primary metabolism, so these data suggest a functional association with metabolic gene control.

There is a possible relationship between MAEB and cellular response to abundant glycine. MAEB is frequently associated with *gcvP* and *gcvT*, which are part of the glycine cleavage system, wherein excess glycine feeds into the citric acid cycle. MAEB is also associated with several citric acid cycle genes (see Supplementary Data). However, MAEB is associated with some other genes with a more tenuous relationship to glycine or the citric acid cycle. It is tempting to infer a relationship to the glycine cleavage system because the highest number of MAEB repeats are associated with the *gcv* genes in this system. Moreover, there exists at least one riboswitch class that binds glycine (42), but this class is present in only one copy per genome in organisms with MAEB, possibly leaving a role in glycine regulation for MAEB.

Representatives of the MAEB motif exhibit covariation that preserves base pairing, but others carry mutations that disrupt pairing. This fact suggests that it could, in fact, be a DNA-sequence that binds a protein dimer, such that each protein unit binds to opposite strands. However, one characteristic is inconsistent with this hypothesis. Nucleotides at one pair of symmetric positions are conserved as purines (A or G) in both sides of the stem. Since purines are never Watson–Crick pairs, they could not have the same identity on opposite strands. Although it is expected that instances of a DNA-binding motif will differ, the symmetric purines imply that the motif itself (and not merely the instances) has a distinct pattern on opposite strands.

Although we cannot rule out the possibility that MAEB could be a repetitive element, its association with metabolic genes argues against this hypothesis. It is also possible that MAEB is part of a protein-binding RNA like CsrB. CsrB is an RNA with roughly 18 hairpins, each of which can bind one CsrA protein subunit (55).

The mini-*ykkC* motif

The mini-*ykkC* motif consists of two tandem hairpins whose stems show considerable covariation and whose loops have characteristic ACGR motifs (Figure 1). Mini-*ykkC* is widespread in α-, β- and γ-proteobacteria, with additional examples in other taxa. We named this motif mini-*ykkC* because it appears to be a *cis*-regulator of a set of genes similar to that of the previously described *ykkC/yxkD motif* (7) (hereafter termed '*ykkC*'). The Supplementary Data lists all eight conserved domains common to *ykkC* and mini-*ykkC*. However, the structures of *ykkC* and mini-*ykkC* appear to be unrelated.

The *ykkC* motif is a highly structured and broadly conserved motif that was proposed previously to be a promising riboswitch candidate. The simple structure of mini-*ykkC*, however, is uncharacteristic of most other riboswitches, though its broad phylogeny suggests a function that dictates broad conservation. Mini-*ykkC* appears to be a *cis*-regulatory element because it is associated with a relatively narrow set of gene functions and it is near to their coding sequences (90% are within 33 nt of the Shine–Dalgarno sequence). We propose that mini-*ykkC* serves the same (but currently unknown) role as the *ykkC* motif, although the mechanisms used to control gene expression could be different.

We note that there might be instances of mini-*ykkC* with only one hairpin. However, we did not explore this possibility because the simplicity of the single hairpin would lead to a prohibitively high false positive rate in genome-scale searches. This issue is not a problem for the full, two-hairpin motif (see Supplementary Data).

The *purD* motif

The *purD* motif is found upstream of all *purD* genes in fully sequenced ε-proteobacteria (e.g. *Campylobacter* and *Helicobacter*). The *purD* gene encodes GAR (phosphoribosylglycinamide) synthetase, which is involved in purine biosynthesis. The *purD* motif shows covariation and modular stems, although it also exhibits some mutations that disrupt base pairing (Figure 1).

To test the hypothesis that the *purD* motif represents a riboswitch aptamer, we used in-line probing assays (56) to test for binding of the RNA against a panel of available purine compounds, including GAR (see Supplementary Data for list). Our assays showed no evidence of structural modulation induced by any of these compounds (data not shown). Although these data fail to support the hypothesis that the *purD* motif is a riboswitch, its consistent association with the *purD* gene at least implies a *cis*-regulatory role.

The 6C motif

This motif is widespread among Actinobacteria, and consists of two hairpins, where the loop of each contains a run of at least 6 Cs. The 6C motif exhibits significant covariation in its stems. 6C motif instances are usually moderately close (200–300 nt) to genes predicted to be related to chromosome partitioning and pilus assembly.

However, given its distance, it is not clear whether 6C is functionally related to these genes.

Transposon- and excisionase-associated motifs

We also found one transposase-associated motif in α-proteobacteria and another associated with Xis excisionases in Actinobacteria, although the excisionase genes could be misclassified (see Supplementary Data). Both motifs consist primarily of a single hairpin with a 10-to-15-bp stem. Both motifs also exhibit much covariation, which suggests they form functional, structured RNAs. The excisionase motif is in a 5′ regulatory configuration to the excisionase gene.

However, both motifs could also be dsDNA sites recognized by protein dimers, where each subunit binds to sites on opposite strands. Alternately, either motif could conceivably function as a structural dsDNA element. A hairpin element, combined with other factors, could favor a structure with two intra-strand hairpins embedded in dsDNA, a 'cruciform'-like structure that is the preferred target for proteins in distinct, though related contexts (57,58). DNA-binding motifs in Xis were also described (59), although no motifs containing hairpins were reported.

The ATPC motif

The ATPC motif occurs in some Cyanobacteria in an ATP synthase operon, between genes encoding the A and C subunits. ATPC motif instances are found in all sequenced strains of *Prochlorococcus marinus*, and certain species of *Synechococcus*. The motif consists primarily of a three-stem junction. Previous studies have proposed hairpin-like structures in Cyanobacterial ATP synthase operons, but not more complicated shapes, and in different locations from the ATPC motif (60).

The cyano-30S motif

This motif occurs in some Cyanobacteria and is in a 5′ regulatory configuration to genes encoding 30S ribosomal protein S1. It consists of two hairpins that are dissimilar to each other, and a pseudoknot wherein five nucleotides in the P1 loop base-pair with nucleotides just beyond the 3′ end of P2. Although there are several mutations that disrupt pairing in P1 and P2, there are also many compensatory mutations in these stems. Moreover, the pairing in the pseudoknot covaries and is present in all representatives.

Given the gene context, we expect that this motif mimics the ligand of the downstream ribosomal protein gene (54), and that the product of this gene thereby controls its own expression. Although we commented on ribosomal protein gene autoregulation previously in Firmicutes (5), we generally ignored ribosomal-gene-associated RNA motifs in the present study because many have already been characterized. However, the identification of the cyano-30S motif supports the view that such RNAs are found in a wide variety of phyla.

Lactobacillales motifs

The lacto-1 and lacto-2 motifs are confined to the order Lactobacillales. The lacto-1 motif has some covariation, but some mutations disrupt base pairing, so its assignment as a structured RNA is uncertain. Some instances intersect a variable region of the S(MK) (or SAM-III) (61) riboswitch between the main hairpin and the Shine–Dalgarno sequence. The lacto-2 motif consists of a large hairpin with many internal loops, some of which have highly conserved sequences. Although some mutations disrupt pairing, there is a considerable amount of covariation, which suggests that the lacto-2 motif instances are probably structured RNAs.

TD (Treponema denticola) motifs

Two predicted motifs have several instances in *Treponema denticola*, but are not found in any other sequenced bacteria. The motifs, TD-1 and TD-2 have 28 and 36 representatives, respectively. Seven TD-1 motif representatives overlap reverse complements of instances of TD-2, and share the two 5′-most hairpins. Although it is possible that the two motifs could be merged, it is not obvious how, because there is significant variation in the non-overlapping instances.

Both motifs show covariation and either variable-length or modular stems. However, the modest but noticeable number of mutations that disrupt pairing reduces confidence that they are functional RNAs. The TD-1 motif is usually in a 5′ regulatory configuration to genes, although the wide array of poorly characterized genes makes it difficult to suggest a coherent *cis*-regulatory function. The TD-2 motif does not share the 5′ regulatory configuration, so it could correspond to a non-coding RNA.

DISCUSSION

Using the CMfinder-based comparative genomics pipeline, we found 22 novel putative RNA motifs. Two have already been experimentally confirmed as riboswitches. For several others, covariation and other characteristics suggest that they are functional structured RNAs, and we have proposed possible functions for many of the motifs. Thus, our pipeline appears to be useful for discovering novel RNAs, which in turn will contribute to our understanding of RNA biochemistry and bacterial gene regulation.

Our findings here and previously (5) demonstrate that the CMfinder-based pipeline is usually able to recover RNAs that are widespread, possess a highly conserved and extensive secondary structure, are roughly 60 nt or more in length, and are associated with homologous genes. Three candidate riboswitches have these characteristics (GEMM, Moco and SAH). The remaining three candidates, SAM-IV, the COG4708 motif and the *sucA* motif are more narrowly distributed than most known riboswitches in that none of these motifs is found outside a single order in taxonomy level. This observation suggests that many of the undiscovered riboswitch classes have more narrow phylogenetic distributions than those discovered previously.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Nicholas Carriero and Robert Bjornson for technical assistance and members of the Breaker laboratory for helpful discussions. This work was funded by grants from the NIH (GM 068819, R01 HG02602) and NSF (EIA-0323510, DBI-0218798). J.E.B. was a Howard Hughes Medical Institute predoctoral fellow. RNA science in the Breaker laboratory is also supported by the HHMI. This work extensively used computers from the Yale Life Sciences High Performance Computing Center (NIH grant RR19895-02). Funding to pay the Open Access publication charges for this article was provided by HHMI.

Conflict of interest statement. None declared.

REFERENCES

1. He,L. and Hannon,G.J. (2004) microRNAs: small RNAs with a big role in gene regulation. *Nat Rev. Genet.*, **5**, 522–531.
2. Kima,V.N. and Nam,J.-W. (2006) Genomics of microRNA. *Trends Genet.*, **22**, 165–173.
3. Storz,G., Altuvia,S. and Wassarman,K.M. (2005) An abundance of RNA regulators. *Annu. Rev. Biochem.*, **74**, 199–217.
4. Claverie,J.M. (2005) Fewer genes, more noncoding RNA. *Science*, **309**, 1529–1530.
5. Yao,Z., Barrick,J.E., Weinberg,Z., Neph,S., Breaker,R.R., Tompa,M. and Ruzzo,W.L. (2007) A computational pipeline for high throughput discovery of *cis*-regulatory noncoding RNA in prokaryotes. *PLoS Comput. Biol.*, **3**, e126.
6. Yao,Z., Weinberg,Z. and Ruzzo,W.L. (2006) CMfinder—a covariance model based RNA motif finding algorithm. *Bioinformatics*, **22**, 445–452.
7. Barrick,J.E., Corbino,K.A., Winkler,W.C., Nahvi,A., Mandal,M., Collins,J., Lee,M., Roth,A., Sudarsan,N. *et al.* (2004) New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proc. Natl Acad. Sci. USA*, **101**, 6421–6426.
8. Corbino,K.A., Barrick,J.E., Lim,J., Welz,R., Tucker,B.J., Puskarz,I., Mandal,M., Rudnick,N.D. and Breaker,R.R. (2005) Evidence for a second class of *S*-adenosylmethionine riboswitches and other regulatory RNA motifs in alpha-proteobacteria. *Genome Biol.*, **6**, R70.
9. Axmann,I.M., Kensche,P., Vogel,J., Kohl,S., Herzel,H. and Hess,W.R. (2005) Identification of cyanobacterial non-coding RNAs by comparative genome analysis. *Genome Biol.*, **6**, R73.
10. Seliverstov,A.V., Putzer,H., Gelfand,M.S. and Lyubetsky,V.A. (2005) Comparative analysis of RNA regulatory elements of amino acid metabolism genes in Actinobacteria. *BMC Microbiol.*, **5**, 54.
11. McCutcheon,J.P. and Eddy,S.R. (2003) Computational identification of non-coding RNAs in *Saccharomyces cerevisiae* by comparative genomics. *Nucleic Acids Res.*, **31**, 4119–4128.
12. Coventry,A., Kleitman,D.J. and Berger,B. (2004) MSARI: multiple sequence alignments for statistical detection of RNA secondary structure. *Proc. Natl Acad. Sci. USA*, **101**, 12102–12107.
13. Washietl,S., Hofacker,I.L., Lukasser,M., Hüttenhofer,A. and Stadler,P.F. (2005) Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat. Biotechnol.*, **23**, 1383–1390.
14. Pedersen,J.S., Bejerano,G., Siepel,A., Rosenbloom,K., Lindblad-Toh,K., Lander,E.S., Kent,J., Miller,W. and Haussler,D. (2006) Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.*, **2**, e33.
15. Torarinsson,E., Sawera,M., Havgaard,J.H., Fredholm,M. and Gorodkin,J. (2006) Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Res.*, **16**, 885–889.
16. Winkler,W.C. and Breaker,R.R. (2005) Regulation of bacterial gene expression by riboswitches. *Annu. Rev. Microbiol.*, **59**, 487–517.
17. Batey,R.T. (2006) Structures of regulatory elements in mRNAs. *Curr. Opin. Struct. Biol.*, **16**, 299–306.
18. Marchler-Bauer,A., Anderson,J.B., Cherukuri,P.F., DeWeese-Scott,C., Geer,L.Y., Gwadz,M., He,S., Hurwitz,D.I., Jackson,J.D. *et al.* (2005) CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Res.*, **33**, 192–196.
19. Pruitt,K., Tatusova,T. and Maglott,D. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **33**, 501–504.
20. Neph,S. and Tompa,M. (2006) MicroFootPrinter: a tool for phylogenetic footprinting in Prokaryotic Genomes. *Nucleic Acids Res.*, **34**, W366–W368.
21. Weinberg,Z. and Ruzzo,W.L. (2004) *RECOMB04: Proceedings of the Eighth Annual International Conference on Computational Molecular Biology*, Faster genome annotation of non-coding RNA families without loss of accuracy San Diego, CA, pp. 243–251.
22. Weinberg,Z. and Ruzzo,W.L. (2004) Exploiting conserved structure for faster annotation of non-coding RNAs without loss of accuracy. *Bioinformatics*, **20**, i334–i341.
23. Weinberg,Z. and Ruzzo,W.L. (2006) Sequence-based heuristics for faster annotation of non-coding RNA families. *Bioinformatics*, **22**, 35–39.
24. Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.*, **22**, 2079–2088.
25. Eddy,S.R. (2005) *Infernal User's Guide*. ftp://ftp.genetics.wustl.edu/pub/eddy/software/infernal/Userguide.pdf
26. Klein,R.J. and Eddy,S.R. (2003) RSEARCH: finding homologs of single structured RNA sequences. *BMC Bioinformatics*, **4**, 44.
27. Griffiths-Jones,S., Moxon,S., Marshall,M., Khanna,A., Eddy,S.R. and Bateman,A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, 121–124.
28. Griffiths-Jones,S. (2005) RALEE-RNA ALignment Editor in Emacs. *Bioinformatics*, **21**, 257–259.
29. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
30. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
31. Wan,X.-F. and Xu,D. (2004) Intrinsic terminator prediction and its application in *Synechococcus* sp. WH8102. *J. Comp. Sci. Tech.*, **20**, 465–482.
32. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–357.
33. Tyson,G.W., Chapman,J., Hugenholtz,P., Allen,E.E., Ram,R.J., Richardson,P.M., Solovyev,V.V., Rubin,E.M., Rokhsar,D.S. *et al.* (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, **428**, 37–43.
34. Venter,J.C., Remington,K., Heidelberg,J.F., Halpern,A.L., Rusch,D., Eisen,J.A., Wu,D., Paulsen,I., Nelson,K.E. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.
35. Rusch,D.B., Halpern,A.L., Sutton,G., Heidelberg,K.B., Williamson,S., Yooseph,S., Wu,D., Eisen,J.A., Hoffman,J.M. *et al.* (2007) The Sorcerer II global ocean sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol.*, **5**, e77.
36. Gerstein,M., Sonnhammer,E.L.L. and Chothia,C. (1994) Volume changes in protein evolution. *J. Mol. Biol.*, **236**, 1067–1078.
37. Henkin,T.M. and Yanofsky,C. (2002) Regulation by transcription attenuation in bacteria: how RNA provides instructions for transcription termination/antitermination decisions. *Bioessays*, **24**, 700–707.

38. Hendrix,D.K., Brenner,S.E. and Holbrook,S.R. (2005) RNA structural motifs: building blocks of a modular biomolecule. *Q. Rev. Biophys.*, **38**, 221–243.
39. Costa,M. and Michel,F. (1997) Rules for RNA recognition of GNRA tetraloops deduced by *in vitro* selection: comparison with *in vivo* evolution. *EMBO J.*, **16**, 3289–3302.
40. Sudarsan,N., Hammond,M.C., Block,K.F., Welz,R., Barrick,J.E., Roth,A. and Breaker,R.R. (2006) Tandem riboswitch architectures exhibit complex gene control functions. *Science*, **314**, 300–304.
41. Welz,R. and Breaker,R.R. (2007) Ligand binding and gene control characteristics of tandem riboswitches in *Bacillus anthracis*. *RNA*, **13**, 573–582.
42. Mandal,M., Lee,M., Barrick,J.E., Weinberg,Z., Emilsson,G.M., Ruzzo,W.L. and Breaker,R.R. (2004) A glycine-dependent riboswitch that uses cooperative binding to control gene expression. *Science*, **306**, 275–279.
43. Meibom,K.L., Li,X.B., Nielsen,A.T., Wu,C.-Y., Roseman,S. and Schoolnik,G.K. (2004) The *Vibrio cholerae* chitin utilization program. *PNAS*, **101**, 2524–2529.
44. Kirn,T.J., Jude,B.A. and Taylor,R.K. (2005) A colonization factor links *Vibrio cholerae* environmental survival and human infection. *Nature*, **438**, 863–866.
45. Meibom,K.L., Blokesch,M., Dolganov,N.A., Wu,C.-Y. and Schoolnik,G.K. (2005) Chitin induces natural competence in *Vibrio cholerae*. *Science*, **310**, 1824–1827.
46. Schaffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.
47. Methé,B.A., Nelson,K.E., Eisen,J.A., Paulsen,I.T., Nelson,W., Heidelberg,J.F., Wu,D., Wu,M., Ward,N. *et al.* (2003) Genome of *Geobacter sulfurreducens*: metal reduction in subsurface environments. *Science*, **302**, 1967–1969.
48. Reguera,G., McCarthy,K.D., Mehta,T., Nicoll,J.S., Tuominen,M.T. and Lovley,D.R. (2005) Extracellular electron transfer via microbial nanowires. *Nature*, **435**, 1098–1101.
49. Mehta,T., Coppi,M.V., Childers,S.E. and Lovley,D.R. (2005) Outer membrane *c*-type cytochromes required for Fe(III) and Mn(IV) oxide reduction in *Geobacter sulfurreducens*. *Appl. Environ. Microbiol.*, **71**, 8634–8641.
50. Kim,B.-C., Qian,X., Leang,C., Coppi,M.V. and Lovley,D.R. (2006) Two putative *c*-type multiheme cytochromes required for the expression of OmcB, an outer membrane protein essential for optimal Fe(III) reduction in *Geobacter sulfurreducens*. *J. Bact.*, **188**, 3138–3142.
51. Bassler,B.L. and Losick,R. (2006) Bacterially speaking. *Cell*, **125**, 237–246.
52. Ueland,P.M. (1982) Pharmacological and biochemical aspects of *S*-adenosylhomocysteine and *S*-adenosylhomocysteine hydrolase. *Pharmacol. Rev.*, **34**, 223–253.
53. Roth,A., Winkler,W.C., Regulski,E.E., Lee,B.W., Lim,J., Jona,I., Barrick,J.E., Ritwik,A., Kim,J.N. *et al.* (2007) A riboswitch selective for the queuosine precursor preQ$_{(1)}$ contains an unusually small aptamer domain. *Nat. Struct. Mol. Biol*, **14**, 308–317.
54. Zengel,J.M. and Lindahl,L. (1994) Diverse mechanisms for regulating ribosomal protein synthesis in *Escherichia coli*. *Prog. Nucleic Acid Res. Mol. Biol.*, **47**, 331–370.
55. Liu,M.Y., Gui,G., Wei,B., Preston,J.F.III, Oakford,L., Yüksel,Ü., Giedroc,D.P. and Romeo,T. (1997) The RNA molecule CsrB binds to the global regulatory protein CsrA and antagonizes its activity in *Escherichia coli*. *J. Biol. Chem.*, **272**, 17502–17510.
56. Soukup,G.A. and Breaker,R.R. (1999) Relationship between internucleotide linkage geometry and the stability of RNA. *RNA*, **5**, 1308–1325.
57. Potamana,V.N., Shlyakhtenkob,L.S., Oussatchevaa,E.A., Lyubchenkob,Y.L. and Soldatenkov,V.A. (2005) Specific binding of Poly(ADP-ribose) polymerase-1 to cruciform hairpins. *J. Mol. Biol.*, **348**, 609–615.
58. Posey,J.E., Pytlos,M.J., Sinden,R.R. and Roth,D.B. (2006) Target DNA structure plays a critical role in RAG transposition. *PLoS Biol.*, **4**, 350.
59. Gottfried,P., Kolot,M. and Yagil,E. (2001) The effect of mutations in the Xis-binding sites on site-specific recombination in coliphage HK022. *Mol. Genet. Genomics*, **266**, 584–590.
60. Curtis,S.E. (1988) Structure, organization and expression of cyanobacterial ATP synthase genes. *Photosynth. Res.*, **18**, 223–244.
61. Fuchs,R.T., Grundy,F.J. and Henkin,T.M. (2006) The S(MK) box is a new SAM-binding RNA for translational regulation of SAM synthetase. *Nat. Struct. Mol. Biol.*, **13**, 226–233.

An mRNA structure that controls gene expression by binding FMN

Wade C. Winkler, Smadar Cohen-Chalamish, and Ronald R. Breaker†

Department of Molecular, Cellular, and Developmental Biology, Yale University, P.O. Box 208103, New Haven, CT 06520-8103

Communicated by Gerald F. Joyce, The Scripps Research Institute, La Jolla, CA, October 16, 2002 (received for review September 27, 2002)

The *RFN* element is a highly conserved domain that is found frequently in the 5′-untranslated regions of prokaryotic mRNAs that encode for flavin mononucleotide (FMN) biosynthesis and transport proteins. We report that this domain serves as the receptor for a metabolite-dependent riboswitch that directly binds FMN in the absence of proteins. Our results also indicate that in *Bacillus subtilis*, the riboswitch most likely controls gene expression by causing premature transcription termination of the *ribDEAHT* operon and precluding access to the ribosome-binding site of *ypaA* mRNA. Sequence and structural analyses indicate that the *RFN* element is a natural FMN-binding aptamer, the allosteric character of which is harnessed to control gene expression.

The expression of certain genes is controlled by mRNA elements that form receptors for target metabolites. Selective binding of a metabolite by such an mRNA "riboswitch" permits a shift in the conformation to an alternative structure that results ultimately in the modulation of protein synthesis. For example, the *btuB* gene of *Escherichia coli* carries a highly conserved sequence element termed the B_{12} box (1) in the 5′ UTR of the mRNA that directly binds coenzyme B_{12} with high selectivity (2). This binding event seems to operate via an allosteric mechanism that represses expression of a reporter gene to ≈1% of that observed in cells grown in the absence of added coenzyme B_{12}. Similarly, many organisms carry a highly conserved "*thi* box" sequence (3) in mRNAs that are required for the biosynthesis of the coenzyme thiamine pyrophosphate. An expanded region of RNA that encompasses the *thi* box also has been shown (4) to function as a riboswitch by directly binding thiamine pyrophosphate, resulting in reduced translation of *thiM* and *thiC* mRNAs in *E. coli*. It is conceivable that riboswitches also could modulate transcription termination, although only proteins (5, 6) or RNAs are currently known to regulate transcription termination in trans (7).

Another highly conserved but distinct RNA domain, termed the *RFN* element, has been identified in mRNAs of prokaryotic genes required for the biosynthesis of riboflavin and FMN (8, 9). It is known (10, 11) that FMN is required for down-regulation of the *ribDEAHT* operon (hereafter termed *ribD*) of *Bacillus subtilis*, which encodes several FMN biosynthetic enzymes. Furthermore, mutations within the *RFN* element of *ribD* eliminate FMN-mediated regulation (12, 13). Sequence comparisons of *RFN* elements from various riboflavin biosynthesis genes and *ypaA* (a putative riboflavin transport protein) (8, 14) have been used to generate a secondary structure model (Fig. 1*A*) and to establish the conserved nucleotides of this domain (8, 9). The structural model is composed of a six-stem junction wherein extensive sequence conservation exists at the bases of the stem elements and among the intervening nucleotides that form the core of the junction (Fig. 1*A*).

It has been proposed that either an unidentified FMN-dependent protein effector (15) or perhaps FMN alone (8, 9) binds to the *RFN* element to repress *ribD* expression in *B. subtilis*. To explore the possibility that the *RFN* element serves as a component of an FMN-dependent riboswitch, we prepared RNAs corresponding to the 5′-UTR sequences of *B. subtilis ribD* and *ypaA* mRNAs by *in vitro* transcription. The RNAs were found to undergo FMN-dependent structural alterations in a protein-free mixture. Initial analyses suggest that FMN binding by the *ribD* leader causes transcription termination, whereas FMN binding by the *ypaA* leader results in sequestration of an adjacent ribosome-binding site. These results suggest that the *RFN* element serves as a component of a metabolite-dependent riboswitch that can bind FMN in the absence of proteins. In addition, FMN binding is likely to be responsible for down-regulating the expression of *ribD* and *ypaA* via distinct genetic control mechanisms.

Materials and Methods

Oligonucleotides and Chemicals. Synthetic DNAs were purchased from The Keck Foundation Biotechnology Resource Center at Yale University, purified by denaturing PAGE, and eluted from the gel by crush-soaking in 10 mM Tris·HCl (pH 7.5 at 23°C)/200 mM NaCl/1 mM EDTA. The DNA was recovered from the crush-soak solution by precipitation with ethanol. FMN, FAD, and riboflavin were acquired from Sigma. The radiolabeled nucleotides [α-^{32}P]UTP and [γ-^{32}P]ATP were purchased from Amersham Pharmacia.

Cloning of *B. subtilis ribD* and *B. subtilis ypaA* Regulatory Regions. The nucleotide sequence from −61 to 302 of the *B. subtilis ribDEAHT* operon (16, 17) was amplified by PCR from *B. subtilis* strain 168 (a gift from D. Söll, Yale University) as an *Eco*RI–*Bam*HI fragment. The DNA was ligated into *Eco*RI-, *Bam*HI-digested pGEM4Z (Promega) to generate pGEM4Z-*ribD*. Similarly, the sequence spanning nucleotides −94 to 347 of *B. subtilis ypaA* was used to generate pGEM4Z-*ypaA*. The plasmids were transformed into *E. coli* Top10 cells (Invitrogen) for all other manipulations. All sequences were verified by DNA sequencing (United States Biochemical Thermosequenase).

***In Vitro* Transcription.** DNA templates for preparative *in vitro* transcription of *ribD* and *ypaA* RNAs were produced by PCR amplification of the corresponding regions of the pGEM4Z-*ribD* and pGEM4Z-*ypaA* plasmids, respectively, with the appropriate DNA primers. The DNA primers were designed to replace the endogenous *B. subtilis* promoter sequence with that of a T7 promoter sequence and to introduce a CC dinucleotide into the DNA templates such that both RNAs carry a 5′-terminal GG sequence. RNAs were prepared by *in vitro* transcription as described (17) or through the use of the RiboMAX transcription kit (Promega) according to the manufacturer's directions. The transcription products were resolved by PAGE and examined by PhosphorImager (Molecular Dynamics). RNA products were isolated by denaturing 6% PAGE and 5′ ^{32}P-labeled as described (18).

The addition or deletion of U residues to alter the *ribD* RNA at positions 261–264 was achieved by using mutant oligonucleotides during PCR amplification of the pGEM4Z-*ribD* template DNA. The resulting templates each carry the 20 nucleotides

Abbreviation: SD, Shine–Dalgarno.

†To whom correspondence should be addressed. E-mail: ronald.breaker@yale.edu.

www.pnas.org/cgi/doi/10.1073/pnas.212628899

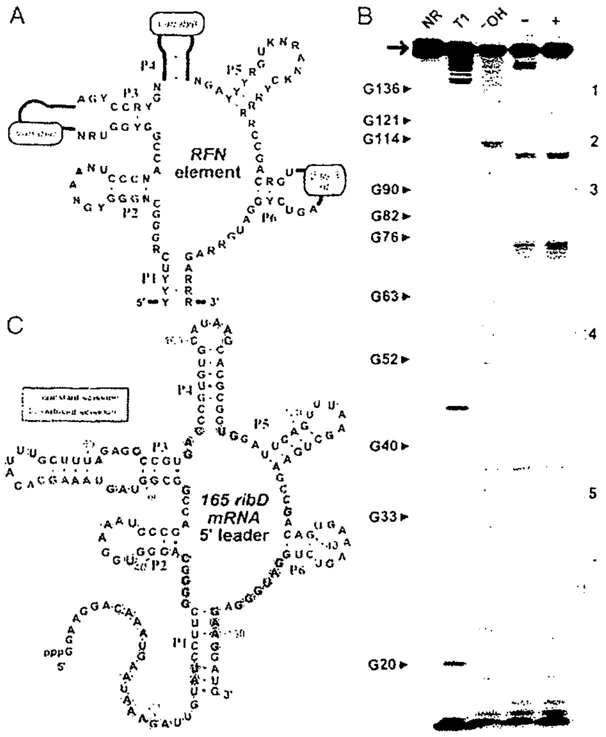


Fig. 1. FMN-induced structure modulation of an *RFN* element. (*A*) Structural model and conserved sequences of the *RFN* element derived from a phylogenetic analysis of prokaryote mRNA sequences. Nucleotides defined by letters are present in >90% of the RNAs examined. The letters R, Y, K, and N represent purine, pyrimidine, G or U, and any nucleotide identity, respectively. The six stem–loop structures are labeled P1–P6. The original model and the sequence data used to create this adaptation are described in ref. 9. (*B*) Structural probing of the 165 *ribD* RNA (arrow) in the presence (+) and absence (−) of 10 μM FMN. Also, 5′ ^{32}P-labeled *ribD* RNA precursors were not reacted (NR) or were subjected to partial digest with RNase T1 (T1) or alkali (-OH) as indicated. Blue arrowheads indicate regions of FMN-dependent modulation of spontaneous cleavage. Regions 1 (nucleotide 136), 2 (nucleotides 113 and 114), 3 (nucleotides 89–91), 4 (nucleotide 56), and 5 (nucleotides 33 and 34) were subsequently used to establish an apparent K_D value for FMN binding (Fig. 2). See *Materials and Methods* for experimental details. (*C*) Secondary-structure model of the *165 ribD* RNA and the nucleotide positions of spontaneous cleavage in the presence (yellow) and absence (yellow and red) of FMN as identified from *B*. Cleavage occurs 3′ relative to the nucleotides highlighted in color.

immediately downstream of position 264 in unaltered form, whereas the poly U-encoding region beginning at position 258 varies from three to seven nucleotides.

In-Line Probing of RNA Constructs. The *B. subtilis* 165 *ribD* and 349 *ypaA* leader mRNAs were subjected to in-line probing (which reveals the structural context of nucleotides based on their relative susceptibility to in-line attack from the adjacent 2′-hydroxyl group) by using a protocol adapted from those described (2, 4, 19, 20). Specifically, ≈1 nM 5′ ^{32}P-labeled RNA was incubated for ≈40 h at 25°C in 20 mM MgCl$_2$/50 mM Tris·HCl (pH 8.3 at 25°C)/100 mM KCl in the presence or absence of added ligand (FMN, FAD, or riboflavin) at concentrations that are indicated for each experiment. All ligand stock solutions and reactions were protected from light by wrapping each tube in aluminum foil. RNA cleavage products of >200 nt in length were resolved by subjecting samples to PAGE for extended time periods (7–21 h) with frequent buffer exchanges. The apparent K_D values for each ligand were established by plotting the normalized fraction of RNA cleaved for each site (relative to the maximum and minimum cleavage values measured) against the logarithm of the concentration of ligand used.

Transcription Termination Assays Using T7 RNA Polymerase. Transcription termination assays were conducted by using RiboMAX transcription kits. Each 10-μl reaction was incubated at 37°C for 2 h and contained 1× buffer (supplied by the manufacturer), 175 μM each NTP, 5 μCi of [α-^{32}P]UTP (1 Ci = 37 GBq), and 0.3 μl of T7 RNAP (supplied by the manufacturer). FMN, riboflavin, or FAD was added to individual transcription reactions as indicated for each experiment. Reaction products were separated by using denaturing 6% PAGE. Bands corresponding to full-length and terminated RNAs were visualized by PhosphorImager, and the yields were established by using IMAGEQUANT software. The fraction of terminated templates was adjusted for the difference in specific activity between the terminated and full-length products. Similarly, transcription termination assays with DNA templates carrying U-insertion or -deletion mutations were incubated as described above for 30 min.

Mapping the Transcription Termination Site. Run-off and FMN-induced termination products for the 304 *ribD* RNA were prepared and separated as described for the transcription termination assays. The resulting RNAs were recovered from the gel by crush-soaking followed by precipitation with ethanol. Both the transcription termination product and run-off transcript were digested separately with a 10–23 deoxyribozyme (21) that was engineered to target *ribD* mRNA for cleavage after position A222. Approximately 40 pmol of RNA were incubated in a 20-μl mixture for 1.5 h at 37°C under conditions similar to those described (22). The 3′ fragments of the deoxyribozyme-cleaved RNAs were separated by denaturing 10% PAGE, recovered from the gel, and 5′ ^{32}P-labeled as described above. The 3′ fragment of the run-off transcript (FL*) was subjected to partial digestion with RNase T1 or alkali as indicated. The resulting RNA fragments were compared with the 3′ fragment of the FMN-induced termination products (T*) by denaturing 10% PAGE and analyzed by PhosphorImager.

Results

The *RFN* Element Binds FMN in the Absence of Protein. To determine whether the *RFN* element serves as part of a metabolite-specific riboswitch, we subjected a portion (nucleotides 1–163) of the *ribD* 5′ UTR of *B. subtilis* to an *in vitro* structure-probing assay (2, 4, 19, 20) in the absence of proteins. This "in-line" probing assay relies on the fact that spontaneous RNA cleavage by internal transesterification typically occurs faster within regions of the RNA chain that are structurally unconstrained, whereas constrained (2°- or 3°-structured) portions of the chain generally undergo spontaneous cleavage less frequently. We find that incubation of 5′ ^{32}P-labeled 165 *ribD* RNA (5′-GG plus the *ribD* 5′-UTR fragment) for 40 h in the presence of 10 μM FMN results in a fragmentation pattern that is distinct from that observed when no effector is added to the incubation (Fig. 1*B*). In contrast, incubation of the 165 *ribD* RNA with 10 μM riboflavin results in a similar but less dramatic change in RNA fragmentation pattern (data not shown), suggesting that riboflavin is bound by the RNA less tightly.

The structure-probing data are consistent also with the secondary structure that has been proposed (8, 9) for the *RFN* element based on phylogenetic sequence comparisons (Fig. 1*A*). For example, the *B. subtilis* 165 *ribD* RNA exhibits fragmentation patterns after incubation in the presence and absence of FMN that are consistent with the formation of stems P1–P5 (Fig. 1*C*),

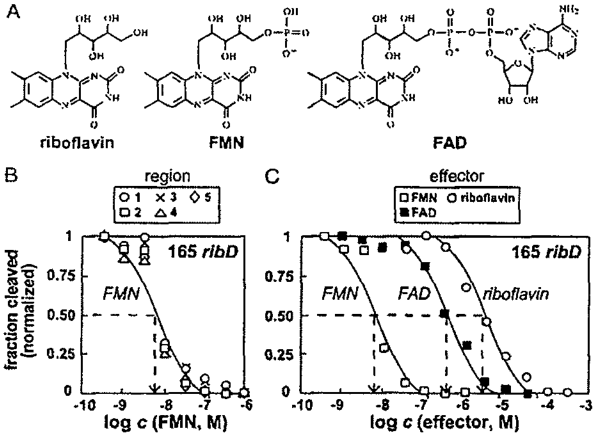


Fig. 2. Affinities and specificity of the 165 *ribD* RNA. (*A*) Chemical structures of riboflavin, FMN, and FAD. (*B*) Determination of the apparent K_D value for FMN binding to the 165 *ribD* RNA. The extent of spontaneous RNA cleavage (normalized relative to the highest and lowest cleavage values measured for each region) was plotted for five regions (Fig. 1*B*) that exhibit FMN-dependent modulation. The dashed line identifies the apparent K_D value or the concentration of FMN required for conversion of half the RNAs into the ligand-bound form. RNAs were subjected to in-line probing as described for Fig. 1 by using various concentrations of ligand as indicated. (*C*) The normalized fraction of spontaneous cleavage at region two for FMN (from *B*) and for the related compounds FAD and riboflavin as indicated. Details are as described for *B*.

because the nucleotides involved in these structural elements exhibit a low level of spontaneous cleavage. In contrast, many of the highly conserved nucleotides that form the internal bulges of the six-stem RNA junction experience a high level of spontaneous cleavage in the absence of FMN but strongly resist fragmentation when FMN is present. These results indicate that the RNA adopts a structural state wherein elements P2–P6 are preorganized and the junction nucleotides adopt a precise FMN-binding fold that is stabilized after ligand association.

The *ribD RFN* Element Discriminates Against Riboflavin and FAD. It has been shown (10) that a disruptive mutation in the riboflavin kinase (FAD synthase) gene causes derepression of riboflavin biosynthetic genes and overproduction of riboflavin. Although FMN and its biologically relevant analogs riboflavin and FAD share the riboflavin moiety (Fig. 2*A*), only FMN and/or FAD serve as the modulating effector(s) for genetic control of the *ribD* RNA. The molecular recognition characteristics of 165 *ribD* RNA were assessed by establishing the apparent K_D value for FMN (Fig. 2*B*) and comparing this value to those determined for riboflavin and FAD (Fig. 2*C*).

The apparent K_D for FMN binding to the 165 *ribD* RNA was established by identifying the concentration of ligand needed to cause ≈50% reduction in the spontaneous cleavage of RNA at each of five regions along the RNA chain. Each of these sites exhibits half-maximal modulation of spontaneous cleavage when ≈5 nM FMN is added to an in-line probing assay. This result is consistent with the hypothesis that the *RFN* element is a tightly binding receptor for FMN and that structural modulation of the RNA is a concerted process that is induced by the presence of ligand.

Interestingly, we find that riboflavin, which differs from FMN by the absence of a single phosphate group (Fig. 2*A*), is bound by the 165 *ribD* RNA less tightly (K_D ≈3 μM) by almost 3 orders of magnitude. This finding is similar to that observed for the thiamine pyrophosphate-specific riboswitch, which discriminates against thiamine and thiamine monophosphate ligands to a similar extent. Thus, the *RFN* element of *ribD* provides another example of an RNA structure that presumably creates productive binding interactions with a negatively charged phosphate moiety despite the fact that RNA itself is polyanionic. These RNAs must form a binding pocket, perhaps by exploiting divalent metal ions, to make productive binding interactions with phosphate.

Furthermore, our in-line probing experiments suggest that FAD is bound by the RNA with an apparent K_D of ≈300 nM. However, we have not eliminated the possibility that our ligand sample might have a small percentage of FMN present, which is expected because of the spontaneous hydrolytic breakdown of the pyrophosphate linkage of FAD. Although our data indicate that the 5′ UTR of *ribD* discriminates between FMN and FAD by at least 60-fold, it is possible that the RNA might achieve a much greater level of discrimination that could be observed only if trace FMN were removed from the FAD sample.

An FMN-Dependent Riboswitch That Causes Transcription Termination. The metabolite-binding domain of a riboswitch must bring about a structural change in the mRNA that modulates gene expression in some defined manner. In a previous study (4) it was shown that a thiamine pyrophosphate-dependent riboswitch from *E. coli thiM* RNA is comprised of an aptamer domain that binds the effector and an "expression platform" that is directly responsible for altering gene expression. We speculate that different riboswitches will make use of expression platforms that provide distinct mechanisms for genetic control and that these expression platforms might have a modular character that can be dissected to establish their mode of operation. Accordingly, we conducted a series of experiments to establish the mechanism of the expression platform for the FMN-dependent riboswitch of *ribD*.

It has been proposed (9) that certain *RFN* elements, including the one present in the *ribD* RNA, might use a transcription termination mechanism for controlling gene expression. Genes from many prokaryotes carry an *RFN* element that is followed by complementary domains that can form transcription terminator or antiterminator stem structures. Thus, the metabolite-dependent control of transcription termination could result if the binding of FMN modulates the formation of these structural elements. In preparation for testing this hypothesis, we created a DNA template that encodes the larger 304-nt leader sequence of the *ribD* RNA (termed 304 *ribD*; Fig. 3*A*). This construct carries the *RFN* element and includes additional nucleotides that encompass the putative terminator and antiterminator elements. Also present in the construct are two domains of six U residues, the first beginning at position 239 and the second beginning at position 258. This second U-rich domain resides immediately 3′ relative to the terminator stem, and together these elements resemble bacterial intrinsic terminators (24, 25). Therefore, this second U-rich domain was identified as the most likely location for transcription termination.

Using an *in vitro* transcription assay based on the action of bacteriophage T7 RNA polymerase, we transcribed the extended DNA template in the absence or presence of 100 μM FMN, FAD, or riboflavin (Fig. 3*B*). Although ≈10% of the transcripts terminate within this second U-rich domain in most instances, the addition of FMN selectively enhances transcription termination to ≈30%. Similar results are obtained when *in vitro* transcription is carried out with *E. coli* RNA polymerase by using a related DNA template (data not shown). Therefore, we believe that T7 RNA polymerase serves as a functional surrogate for a bacterial RNA polymerase in our transcription termination assays. Our data also indicate that binding of FMN to the nascent RNA transcript results in the formation of a structure that promotes transcription termination with unrelated RNA polymerases.

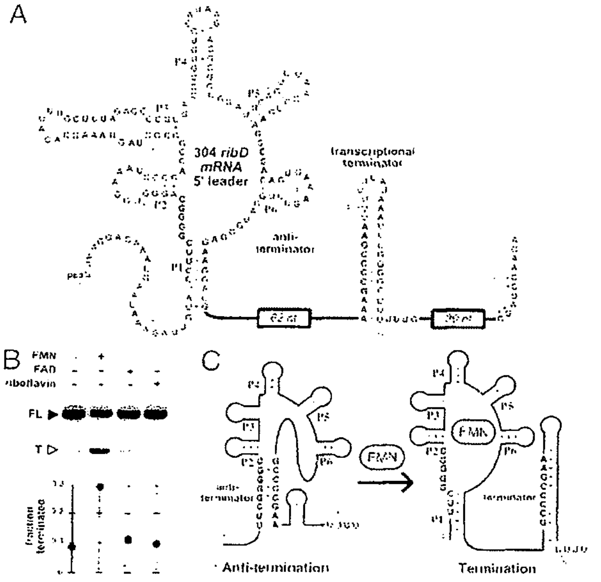


Fig. 3. FMN causes transcription termination of the *ribD* RNA *in vitro*. (*A*) Sequence and secondary structure of the 304 *ribD* RNA. Shaded regions identify nucleotides that are complementary and might serve as an antiterminator structure. Nucleotides denoted with an asterisk have been altered from the wild-type sequence to generate a restriction site. The nucleotide sequence comprising the 62- and 30-nt regions (not shown) are presented in refs. 16 and 17. (*B*) *In vitro* transcription termination assays. *In vitro* transcriptions were conducted by using T7 RNA polymerase and a double-stranded DNA construct that serves as a template for the synthesis of 304 *ribD* RNA. Reactions were incubated in the absence (−) or presence (+) of 100 μM of the compounds as indicated for each lane. FL and T denote full-length and terminated RNA transcripts, respectively. The fraction of total transcripts that are termination products is plotted in the graph. See *Materials and Methods* for experimental details. (*C*) Model for the FMN-dependent riboswitch. Shaded regions identify the putative antiterminator structure that is disrupted after binding of FMN and formation of the P1 structure.

In the 165 *ribD* construct, the nucleotides between P1 and P2 undergo a significant reduction in spontaneous cleavage, which indicates that this region is involved in forming an ordered structure that is necessary for FMN binding. These same nucleotides are predicted also to participate in the formation of the antiterminator structure (9). Therefore, these data are consistent with the proposed riboswitch mechanism (Fig. 3*C*) in which the structures of the effector-bound *RFN* element and the antiterminator stem are mutually exclusive.

Mapping the FMN-Induced Transcription Termination Site. The transcription termination site for the 304 *ribD* RNA construct was identified by comparison of RNAs that correspond to the FMN-induced transcription termination product with the run-off transcript. Because the original transcripts from the 304 *ribD* template are too large to obtain single-nucleotide resolution by PAGE, we used a strategy wherein the transcription products were digested with a deoxyribozyme that catalyzes site-specific cleavage of the RNAs between nucleotides 222 and 223 (Fig. 4*A*). The subsequent 5′ ^{32}P labeling of the 3′-digestion fragments of the full-length and terminated transcripts (to generate FL* and T*, respectively) permitted the high-resolution mapping of the termination site by PAGE (Fig. 4*B*). We find that transcription termination occurs between nucleotides U261 and U263, which are the last U residues in the U-rich domain that immediately follows the terminator stem.

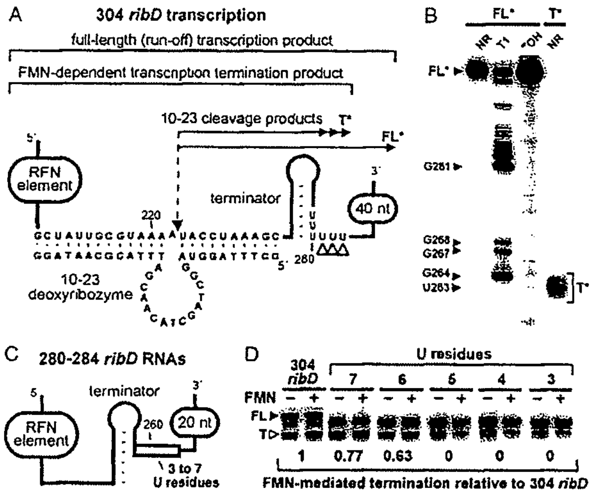


Fig. 4. Mapping of the transcription termination site and the importance of the U-rich domain. (*A*) The 304 *ribD* construct (see also Fig. 3*A*) and processed fragments used to map the transcription termination site. The nucleotides presented depict the interaction with a 10–23 deoxyribozyme and the U-rich region where transcription termination was expected to occur. The dashed arrow identifies the location of deoxyribozyme-mediated cleavage. The portions of the construct that correspond to the full-length product, the FMN-induced transcription product and the deoxyribozyme digestion products FL* and T* that are derived from the full-length and terminated RNAs, respectively, are also identified. The open arrowheads identify the site of FMN-modulated transcription termination. (*B*) PAGE analysis of the 5′ ^{32}P-labeled FL* and T* RNAs. RNAs were subjected to no additional reaction (NR) or were subjected to partial digestion with RNase T1 (T1) or alkali (-OH) as indicated for each lane. Bands corresponding to the FL* and T* RNAs, along with several T1 digestion products, are identified. The T* RNA corresponds to termination at nucleotides U261–U263, the mobility of which differs from the markers by 1 nt equivalent due to the absence of a 2′,3′-cyclic phosphate (unpublished data). (*C*) Schematic representation of *ribD* RNAs ranging from 280 to 284 nucleotides that carry three through seven U residues, respectively, in the U-rich domain (shaded box). Other than the U insertion or deletions, the RNA is identical in sequence to that of nucleotide 1–284 of the 304 *ribD* RNA. (*D*) Transcription termination assays with templates that encode the 304 *ribD* RNA and for 280–284 ribD variants that carry three through seven U residues, respectively, as indicated for each lane. Transcription reactions were conducted as described for Fig. 3*B* and incubated in the absence (−) or presence (+) of 100 μM FMN. FL and T denote full-length and terminated RNA transcripts, respectively. The numbers below each gel image reflect the relative level of FMN-induced termination compared with the nonmutated 304 *ribD* construct.

To explore the importance of the second U-rich domain in greater detail, we constructed a series of *ribD* constructs in which the number of U residues comprising this domain were varied from three to seven (Fig. 4*C*). *In vitro* transcription of the DNA templates that encode either seven or six U residues yielded near wild-type levels of FMN-induced transcription termination (Fig. 4*D*). In sharp contrast, constructs that carry fewer than six U residues do not exhibit FMN-dependent modulation of transcription termination. These observations indicate that transcription termination occurs at positions near the end of the second U-rich domain and that a stretch of at least six U residues is required for FMN-modulated termination.

The *ypaA* mRNA Riboswitch Uses a Shine–Dalgarno (SD) Sequestration Mechanism. Comparative sequence analysis (9) indicates that a second type of expression platform also can be brought under the control of *RFN*-based riboswitches. For example, the *ypaA* gene, which encodes a riboflavin transport protein, carries an *RFN*

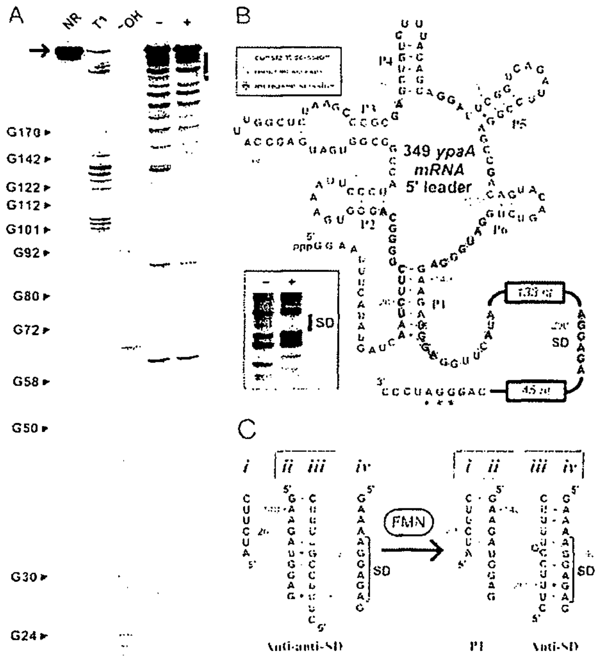


Fig. 5. FMN-induced structure modulation of the *ypaA* riboswitch. (*A*) Structural probing of the 349 *ypaA* RNA (arrow) in the presence (+) and absence (−) of 100 μM FMN. The bar identifies the region of cleavage products that is expanded in *B Inset*. Additional details are as described for Fig. 1*B*. (*B*) Secondary-structure model of the 349 *ypaA* RNA and the nucleotide positions of spontaneous cleavage in the presence (yellow and green) and absence (yellow and red) of FMN as identified from *A*. Nucleotides marked with an asterisk have been altered from the wild-type sequence to generate a restriction site. Sequences not depicted in the 133- and 45-nt domains can be obtained from the *B. subtilis* genome sequence (23). (*Inset*) An image of an extended PAGE separation of the region encompassing the SD element. (*C*) Proposed mechanism for the FMN-modulated formation of structures within the expression platform that control translation. RNA domains *ii* and *iii* form in the absence of FMN, thus exposing the SD element for ribosome binding. In contrast, FMN binding requires the formation of P1, which occupies RNA domain *ii* and permits the sequestration of the SD element by RNA domain *iii*.

element (Fig. 5) that is followed by putative anti-SD and anti-anti-SD elements. In-line probing of a 349-nt construct that encompasses the 5′ UTR of *ypaA* (Fig. 5*A*) demonstrates that the RNA undergoes significant structural modulation after the addition of FMN. Nearly all structure modulation within the 349 *ypaA* RNA (Fig. 5*B*) occurs at positions corresponding to those that undergo FMN-induced change in *ribD* RNA.

In addition, we noticed that the 349 *ypaA* RNA exhibits structural modulation in the region occupied by the SD element (Fig. 5*B Inset*). This observation indicates that the expression platform indeed might be exploiting the FMN-dependent formation of a structure that restricts ribosome access to the SD element. A model for the mechanism of this expression platform is depicted in Fig. 5*C*. In the absence of FMN, the anti-anti-SD element (domain *ii*) is free to base-pair with the anti-SD element (domain *iii*), thus presenting the SD element (domain *iv*) in an unencumbered structural context. However, FMN-binding to the *RFN* element requires formation of the P1 stem between domains *i* and *ii*, which permits the anti-SD to pair with and sequester the SD element. Consistent with this model is the fact that the 5′-most nucleotides (139–143) within domain *ii* do not undergo significant modulation of spontaneous cleavage after FMN addition, whereas the 3′-most nucleotides (positions 144–147) exhibit an increase in spontaneous cleavage (Fig. 5*B*). This observation is expected if the 5′ portion of domain *ii* is base-paired in the absence (anti-anti-SD structure) or presence (P1 structure) of FMN, whereas the 3′ portion of domain *ii* is unpaired only when FMN is bound.

Discussion

RNA structural probing studies with the 5′ UTRs of *ribD* and *ypaA* RNAs confirm that the highly conserved *RFN* element is a natural FMN-binding aptamer. This RNA motif exhibits an exceptionally high affinity for its target ligand (apparent K_D of <10 nM). As with the two natural metabolite-binding aptamers reported (2, 4), the FMN-binding domain of *ribD* exhibits a high level of discrimination against closely related compounds. Furthermore, both the thiamine pyrophosphate- and the FMN-dependent aptamers require the presence of phosphate groups on their respective ligands to bind with the highest affinity, which is a somewhat surprising achievement for a polyanionic receptor molecule. These aspects of molecular recognition are of particular importance in a biological setting, where the promiscuous binding of closely related biosynthetic intermediates would interfere with proper regulation of genetic expression.

The role of the natural FMN aptamer in these two instances most likely is to serve as the recognition domain for FMN-dependent riboswitches that control gene expression of the riboflavin biosynthetic operon and the riboflavin transporter in *B. subtilis*. Preliminary investigations into the mechanisms used by the associated expression platforms in both cases are consistent with those proposed from comparative sequence analysis data (9). Specifically, *ribD* RNA undergoes an increased frequency of transcription termination after the addition of FMN. This is perhaps the most efficient riboswitch mechanism for controlling the expression of large operons, because termination of transcription in the 5′ UTR prevents the synthesis of long mRNAs when their translation is unnecessary. Indeed, regulation by transcription termination may be common among many bacterial species (26). A search for sequences in the *B. subtilis* genome that could form terminator–antiterminator elements revealed nearly 200 candidates (27). In contrast, the riboswitch in the *ypaA* mRNA leader seems to control ribosome access to the SD element. Although the entire sequence of this smaller mRNA would be produced, this genetic control mechanism permits the organism to respond rapidly to declining concentrations of FMN.

Our findings provide additional evidence in support of earlier speculation (3, 8, 28–31) that mRNAs have the ability to play an active role in sensing metabolites for the purpose of genetic control. It seems likely that new riboswitches will be discovered that respond to other metabolites and exhibit more diverse mechanisms of genetic control. The riboswitches examined in this study provide examples of two mechanisms for expression platform function for the down-regulation of gene expression. However, we speculate that there will be instances where gene expression might be increased in response to metabolite binding to mRNAs. For example, certain enzymes that make use of the riboswitch effectors coenzyme B_{12} (2), thiamine pyrophosphate (4), and FMN might make use of expression platforms that permit gene activation. If the occurrence of these or other riboswitches extends across the phylogenetic landscape, or if they are used to control the expression of more than just biosynthetic and transport genes, then it is likely that a greater diversity of genetic control mechanisms will be discovered.

We thank members of the R.R.B. laboratory for comments on the manuscript. This work was supported by National Institutes of Health Grant GM 559343, National Science Foundation Grant EIA-0129939, and a fellowship (to R.R.B.) from The David and Lucile Packard Foundation.

1. Lundrigan, M. D., Köster, W. & Kadner, R. J. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 1479–1483.
2. Nahvi, A., Sudarsan, N., Ebert, M. S., Zou, X., Brown, K. L. & Breaker, R. R. (2002) *Chem. Biol.* **9**, 1043–1049.
3. Miranda-Rios, J., Navarro, M. & Soberón, M. (2001) *Proc. Natl. Acad. Sci. USA* **98**, 9736–9741.
4. Winkler, W., Nahvi, A. & Breaker, R. R. (2002) *Nature* **419**, 952–956.
5. Henkin, T. M. (2000) *Curr. Opin. Microbiol.* **3**, 149–153.
6. Stülke, J. (2002) *Arch. Microbiol.* **177**, 433–440.
7. Grundy, F. J., Winkler, W. C. & Henkin, T. M. (2002) *Proc. Natl. Acad. Sci. USA* **99**, 11121–11126.
8. Gelfand, M. S., Mironov, A. A., Jomantas, J., Kozlov, Y. I. & Perumov, D. A. (1999) *Trends Genet.* **15**, 439–442.
9. Vitreschak, A. G., Rodionov, D. A., Mironov, A. A. & Gelfand, M. S. (2002) *Nucleic Acids Res.* **30**, 3141–3151.
10. Mack, M., van Loon, A. P. G. M. & Hohmann, H.-P. (1998) *J. Bacteriol.* **180**, 950–955.
11. Lee, J.-M., Zhang, S., Saha, S., Santa Anna, S., Jiang, C. & Perkins, J. (2001) *J. Bacteriol.* **183**, 7371–7380.
12. Gusarov, I. I., Kreneva, R. A., Podcharniaev, D. A., Iomantas, I. V., Abalakina, E. G., Stoinova, N. V., Perumov, D. A. & Kozlov, I. I. (1997) *Mol. Biol.* **31**, 446–453.
13. Kil, Y. V., Mironov, V. N., Gorishin, I. Y., Kreneva, R. A. & Perumov, D. A. (1992) *Mol. Gen. Genet.* **233**, 483–486.
14. Kreneva, R. A., Gelfand, M. S., Mironov, A. A., Yomantas, J. A., Kozlov, Y. I., Mironov, A. S. & Perumov, D. A. (2000) *Russ. J. Genet. [Transl. of Genetika (Moscow)]* **36**, 972–974.
15. Perkins, J. B. & Pero, J. (2002) in *Bacillus subtilis and Its Closest Relatives: From Genes to Cells*, eds. Sonenshein, A. L., Hoch, J. A. & Losick, R. (Am. Soc. Microbiol., Washington, DC), pp. 271–286.
16. Azevedo, V., Sorokin, A., Ehrlich, S. D. & Serror, P. (1993) *Mol. Microbiol.* **10**, 397–405.
17. Mironov, V. N., Perumov, D. A., Krayev, A. S., Stepanov, A. I. & Skryabin, K. G. (1990) *Mol. Biol.* **24**, 256–260.
18. Seetharaman, S., Zivarts, M., Sudarsan, N. & Breaker, R. R. (2001) *Nat. Biotechnol.* **19**, 336–341.
19. Soukup, G. A. & Breaker, R. R. (1999) *RNA* **5**, 1308–1325.
20. Soukup, G. A., DeRose, E. C., Koizumi, M. & Breaker, R. R. (2001) *RNA* **7**, 524–536.
21. Santoro, S. W. & Joyce, G. F. (1997) *Proc. Natl. Acad. Sci. USA* **94**, 4262–4266.
22. Pyle, A. M., Chu, V. T., Jankowsky, E. & Boudvillain, M. (2000) *Methods Enzymol.* **317**, 140–146.
23. Kunst, F., Ogasawara, N., Moszer, I. Albertini, A. M., Alloni, G., Azevedo, V., Bertero, M. G., Bessières, P., Bolotin, A., Borchert, S., *et. al.* (1997) *Nature* **390**, 249–256.
24. Wilson, K. S. & von Hippel, P. H. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 8793–8797.
25. Gusarov, I. & Nudler, E. (1999) *Mol. Cell* **3**, 495–504.
26. Henkin, T. M. & Yanofsky, C. (2002) *BioEssays* **24**, 700–707.
27. Merino, E. & Yanofsky, C. (2002) in *Bacillus subtilis and Its Closest Relatives: From Genes to Cells*, eds. Sonenshein, A. L., Hoch, J. A. & Losick, R. (Am. Soc. Microbiol., Washington, DC), pp. 323–336.
28. Gold, L., Brown, D., He, Y., Shtatland, T., Singer, B. S. & Wu, Y. (1997) *Proc. Natl. Acad. Sci. USA* **94**, 59–64.
29. Gold, L., Singer, B., He, Y. & Brody, E. (1997) *Curr. Opin. Genet. Dev.* **7**, 848–851.
30. Nou, X. & Kadner, R. J. (2000) *Proc. Natl. Acad. Sci. USA* **97**, 7190–7195.
31. Stormo, G. D. & Ji, Y. (2001) *Proc. Natl. Acad. Sci. USA* **98**, 9465–9467.

BIOCHEMISTRY